Evolutionary Genomics of Grass Powdery Mildew (*Blumeria graminis*)

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Fabrizio Menardo

von/aus

Italien

Promotionskommission

Prof. Dr. Beat Keller (Vorsitz und Leitung der Dissertation)

PD Dr. Thomas Wicker

Prof. Dr. Kentaro Shimizu

Zürich, 2016

Table of Contents

SUMMARY

Grass powdery mildew is one of the most important cereal diseases in the world and it is caused by the fungus *Blumeria graminis* (Ascomycota). *B. graminis* is an obligate biotroph, and depends on a living host to complete its life cycle. *B. graminis* is divided in *formae speciales (ff.spp.)* which can be distinguished by different host specificities. The availability of a reference genome for the two *ff.spp.* infecting barley and wheat (*f.sp. hordei* and *f.sp. tritici*) allows to perform evolutionary genomic studies in *B. graminis*.

The major part of this thesis describes the characterization of the genome of a new form of powdery mildew, *B.g. triticale*, which emerged recently on the crop triticale, a man-made hybrid between wheat and rye. We found that the genomes of different *B.g. triticale* isolates are mosaics composed of sections from rye powdery mildew (*B.g. secalis*) and wheat powdery mildew (*B.g. tritici*) genomes. This pattern is best explained by a hybridization between a *B.g. secalis* and a *B.g. tritici* isolate followed by two back-crosses with *B.g. tritici*. Moreover we found that there were at least two such hybridization events involving two different pairs of isolates. Our data show that the hybrid of the two mildews specialized on two different hosts can infect the hybrid plant species originating from those two hosts.

In a second study we extended the evolutionary analysis to several other *ff. spp.* of *B. graminis*. In addition to the already available genome sequence of barley and wheat powdery mildew we sequenced the genomes of mildew growing on rye, oat, *Dactylis*, *Lolium* and *Poa*. Using phylogenomic analysis we found a general pattern of co-evolution between pathogen and plant which lasted several millions years. In addition, we identified exceptions to this pattern, namely host jump events and the recent radiation of a clade less than 280,000 years ago. We then applied a coalescent-based method of demographic inference and found evidence of horizontal gene flow between lineages belonging to the recently radiated clade.

In a third project, *B. graminis* effector genes were annotated and studied. Effectors are secreted proteins which are used by the pathogen to manipulate host metabolism and to evade the plant immune system. We identfied new effectors and improved the criteria for their identification and classification. We found that some effector families have homologies with known functional domains indicating possible effector functions. In addition we studied the evolution of effector families in *B. graminis* and found that most of them are present in all forms of *B. graminis*, implicating an ancient origin. Finally we found that most effector families underwent a fast evolution by a combination of positive selection and multiple gene duplications and losses.

ZUSAMMENFASSUNG

Grasmehltau ist eine der wichtigsten Getreidekrankheiten der Welt und wird durch den Pilz *Blumeria graminis* (*Ascomycota*) verursacht. *B. graminis* ist obligat biotroph und somit auf den lebenden Wirt angewiesen, um seinen Lebenszyklus zu vollenden. *B. graminis* ist in *formae speciales (ff.spp.)* unterteilt, diese können auf Grund ihrer Wirtsspezifität unterschieden werden. Die Verfügbarkeit der Referenzgenome der zwei *ff.spp.*, welche Gerste und Weizen infizieren (*f.sp. hordei und f.sp. tritici*) erlaubt Genomevolutionsstudien in *B.graminis*.

Der Grossteil dieser Thesis beschreibt die Charakterisierung des Genoms eines neuen Mehltaustamms, *B.g. triticale*, welcher seit einiges Zeit auf der Getreidesorte Triticale beobachtet wird. Triticale ist ein menschengemachter Hybrid aus Weizen und Roggen. Wir beobachteten, dass die Genome verschiedener *B.g. triticale*-Stämme Mosaike aus Teilen des Genoms des Roggenmehltaus (*B.g. secalis*) und des Weizenmehltaus (*B.g. tritici*) sind. Dieses Muster lässt sich am besten durch eine Hybridisierung von *B.g. secalis*- und *B.g. tritici*-Stämmen erklären, gefolgt von zwei Rückkreuzungen mit *B.g. tritici*. Wir fanden zudem, dass mindestens zwei solcher Hybridisierungen mit jeweils unterschiedlichen Stämmen stattgefunden haben. Unsere Daten zeigen auf, dass der Hybrid zweier Mehltaustämme, welche auf verschiedene Wirte spezialisiert sind, den Hybrid dieser beiden Wirtspflanzen infizieren kann.

In einer zweiten Studie weiteten wir die Evolutionsanalysen auf mehrere andere *B. graminis ff.spp*. aus. Zusätzlich zu den bereits vorhandenen Genomsequenzen von Gersten- und Weizenmehltau sequenzierten wir die Genome von Mehltau, welche auf Roggen, Hafer, *Dactylis*, *Lolium* und *Poa* wachsen. Mithilfe phylogenomischer Analysen fanden wir Muster von Koevolution zwischen Pathogen und Pflanze, welche sich über mehrere Millionen Jahre erstreckt. Zusätzlich identifizierten wir Ausnahmen, namentlich Wirtssprünge und die kürzliche Radiation einer Klade vor weniger als 280‘000 Jahren. Daraufhin nutzten wir koaleszenz-basierte Methoden demographischer Inferenz und fanden Belege horizontalen Genflusses zwischen Abstammungslinien der kürzlichen Radiation.

In einem dritten Projekt, annotierten und studierten wir *B. graminis* Effektorgene. Effektoren sind sekretierte Proteine, welche von Pathogenen genutzt werden um den Wirtsmetabolismus zu manipulieren und das Pflanzenimmunsystem zu umgehen. Wir identifizierten neue Effektoren und verbesserten die Kriterien zu ihrer Identifikation und Klassifikation. Wir beobachteten, dass einige Effektorfamilien Homologie zu bekannten funktionellen Proteindomänen zeigen. Zusätzlich, studierten wir die Evolution der *B. graminis* Effektorfamilien und fanden, dass die meisten Familien

schnell, durch eine Kombination aus positiver Selektion und multipler Genduplikation und –verlust, evoluierten.

CHAPTER 1

General Introduction

The grass powdery mildew (*Blumeria graminis*)

Powdery mildews cause some of the most prominent plant diseases world-wide. They belong to the order of Erysiphales (Ascomycota) and infect leaves, fruits, stems and flowers of almost 10,000 angiosperm species (Glaw 2008). The control of these diseases in agriculture is attained with fungicides and the breeding of resistant plant varieties. The most affected crops are Cucurbitaceae (melons, zucchini, pumpkins etc.) grapes, strawberries and cereals. The taxonomy of powdery mildew is constantly changing and very dynamic, up to now more than 800 species have been described (Braun 2011) based on morphological characteristics and sequence data. While most powdery mildew species attack specific host species, some have a very broad host range. The *Erysiphe alphitoide* species complex is an example of such generalist powdery mildews: the primary hosts are different oak species (*Quercus spp.*), however the fungus is able to infect a large set of tropical trees (including mango, rubber trees and cashew) and asexually reproduce on them (Takamatsu et al. 2007).

Grass powdery mildews are pathogens of wild grasses and domesticated cereals (Fig. 1) that belong to a monospecific genus (*Blumeria*). They are considered to be among the most important fungal pathogens because of their economic impact on cereal crops, most importantly wheat and barley, making them a model system to study biotrophic pathogens (Dean et al. 2012). The only species in the genus, *Blumeria graminis*, is divided in several *formae speciales (ff.spp.)* which are defined by their host plant species. *Forma specialis (f.sp.)* is a taxonomic category that refers to a pathogen adapted to a specific host species and that shows minimal (or no) morphological differences to its closest relative at the species level (Schulze-Lefert and Panstruga 2011). Several studies have shown that host specificity is not absolutely strict in *B. graminis* and should rather be considered an adaptation to a particular host (reviewed in Troch et al. 2014). However, extensive infection experiments have been conducted only with cereal infecting forms and gave discordant results. In particular, while Wyand and Brown (2003) found that isolates of *B. graminis* sampled on barley, wheat, rye and oat infected only the plant species on which they were sampled, several older studies

reported that isolates of the *ff. spp. tritici, hordei* and *avenae* were compatible with a wide range of wild grasses (Eshed and Wahl 1970, Sheng et al. 1993 and 1995). These discrepancies could be due to the different fungal isolates and plant accessions used in the different studies or to the environmental conditions of the infection tests. For example a grass powdery mildew form could grow on a non-adapted host when infecting weak plants with debilitated immune system. It has been proposed that only forms attacking domesticated cereals should be classified as *ff.spp.* This was based on a supposed stronger host specialization observed in these forms (Troch et al. 2014). Two studies (Tosa 1988, Tosa and Sakai 1990) investigated the genetic basis of host specialization in *B. graminis*. In these studies two different *ff. spp.* (*tritici* and *agropyri*) were crossed and their progeny used in infection tests. Based on the observed segregation rates the authors concluded that the avirulence of the *f.sp. agropyri* on wheat is determined by two genes, therefore following the "gene-for-gene" model (Flor 1971). These results are in agreement with the model proposed by Schulze-Lefert and Panstruga (2011) which says that:

1) incompatibility between a pathogen and a non-host plant which is evolutionary similar to the host species is determined mostly by immunity triggered by cytoplasmic resistance protein which usually work in a gene-for-gene manner.

2) Incompatibility between a pathogen and a non-host plant distantly related from the primary host is given by plasma membrane receptors which usually recognize molecular patterns common to many pathogens (See last section of this introduction).

B. graminis, like all other mildews, is an obligate biotroph, depending on a living host to complete its life cycle. It develops a special feeding structure, the haustorium, which develops from the hyphae (Fig. 2). The haustorium is formed through an invagination of the plasma membrane of the plant cell and it is thought to be the structure responsible for the nutrient uptake of the fungus from the plant. Another proposed function of the haustorium is the delivery of effectors. These are fungal proteins which interfere with the plant cell system to avoid defense responses and predispose the establishment and maintenance of the haustorium. However, except few effectors for which the molecular target protein in the plant was identified (Zhang et al. 2012, Schmidt et al. 2014, Pennington et al. 2016) the molecular function of effectors in *B. graminis* is unknown.

B. graminis has a sexual and an asexual life cycle. The asexual cycle begins with a haploid conidiospore landing on the leaf and penetrating the epidermal plant cell wall after formation of an appressorium (Zhang et al. 2005). Inside the plant cell, the fungus forms the haustorium and subsequently develops secondary hyphae which can later form secondary haustoria. On the leaf surface the fungus produces new conidiospores which are then further distributed by wind.

Figure 1. Conidia of *Blumeria graminis* on a wheat leaf. Photograph taken by Francis Parlange.

In nature the sexual cycle is triggered by dry weather at the end of summer, when hyphae of opposite mating types fuse and start a short diploid phase. Ascospores (sexual spores) are produced in fruiting bodies called chasmothecia. Chasmothecia can remain dormant during rough weather conditions, allowing the fungus to over-winter or survive long periods of drought. However, asexual conidiospores are also known to survive winter on not harvested plants and winter wheat seedlings ("green bridges", Liu et al. 2012). Asexual fusion of hyphae (anastomosis) was never observed in *B. graminis* and sexual reproduction has been described as the only opportunity of recombination between different powdery mildew strains (Glawe 2008).

Recently the genomes of barley and wheat powdery mildews have been sequenced (Spanu et al. 2010, Wicker et al. 2013). It was found that the genome of *Blumeria* underwent an expansion due to the activity of repetitive elements which account for more than 90% of the genome. At the same they experienced massive gene losses as extreme adaptation to an obligate pathogenic lifestyle. These achievements promoted much of the research on *B. graminis* in the last years in particular regarding effector biology (see Introduction of chapter 3 for a short review). Moreover the availability of a reference genome allowed the identification of avirulence genes in powdery mildew through genetic mapping (Bourras et al. 2015 Praz et al. 2016) and genome wide association analysis (Praz et al. 2016, Lu et al. 2016). All known avirulence genes in B. graminis are

short (less than 135 amino acids) predicted effectors. Three of them encode proteins that have homology with RNAse domains (AvrPm2 in *B.g. tritici*, AvrA13 and AvrA1 in *B.g. hordei*) while the protein encoded by AvrPm3a/f has no homology with known functional domains.

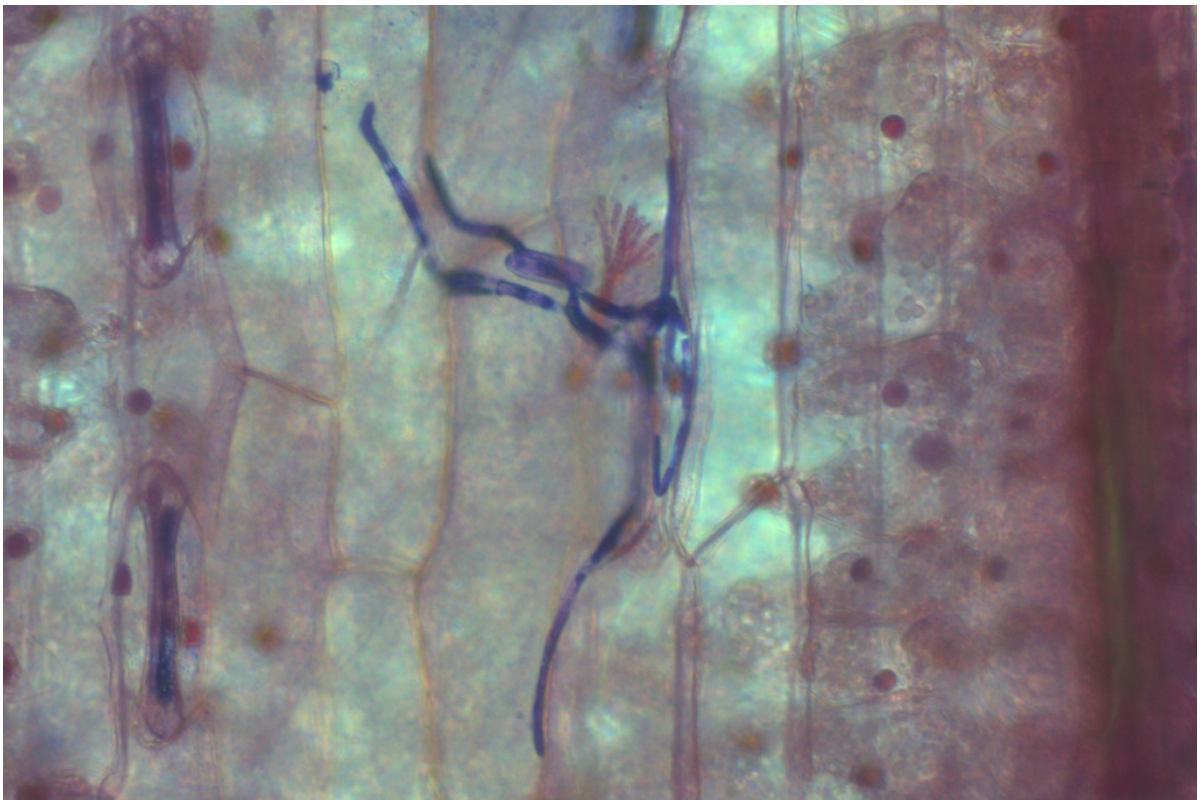


Figure 2. A colony of *B. graminis* on a triticale leaf. The fungal spore and hypae are stained in blue. The bush-like haustorium is stained in red. Photography taken by Marion Müller.

Evolutionary biology in the genomics era

Since the *Origin of species* in 1859 evolutionary biology underwent several developments. Here I present is a subjective and partial summary of major advances with a particular attention to the ones that have occurred in the genomics era.

The rediscovery of the work of Mendel on the inheritance of traits and the development of population genetic theory by Haldane, Fisher and Wright in the beginning of the 20th century provided the genetic basis to the ideas of Darwin on natural selection as driver of evolution. This main conceptual framework for evolutionary biology crystallized in the late fifties in the so called modern synthesis.

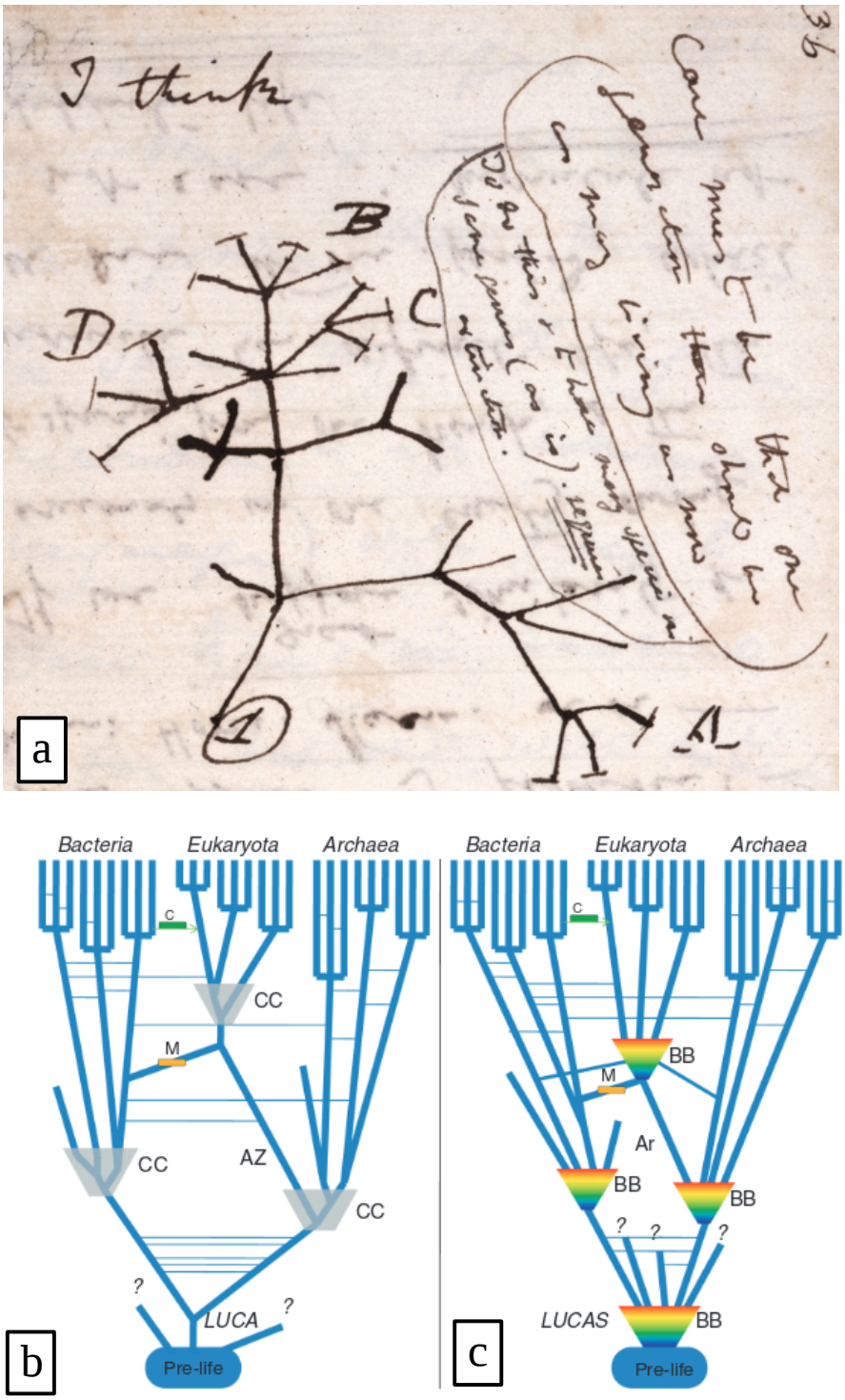


Figure 3. a) The first tree of life (Darwin 1859). **b)** Tree of life with horizontal gene transfers (HGTs), the most massive ones associated with endosymbiosis that originated mitochondria at the base of the Eukaryotic clade and plastids after the split of the main Eukaryotic clades (Koonin 2009). **c)** The Big-bang model. The history of life is represented by tree-like phases with HGT and non-tree-like big-bang (BB) expansions. During these expansion the diversification of lineages is very rapid and involves HGT, recombination and activity of mobile elements (Koonin 2007).

The early population geneticists realized the potential of genetic drift in random fixation of evolutionary changes. However part of the modern synthesis adopted a strict adaptationist view of evolution in which all observed biological diversity originates by random mutation and either gets fixed by natural selection because is beneficial or it is eliminated because disadvantageous (Gould and Lewontin 1979). It is remarkable that this argument is still used and often permeates the thoughts of many non-evolutionary biologists. The main step beyond the adaptationist positions of the modern synthesis was the neutral theory of molecular evolution of Kimura (1983). According to the theory of Kimura most mutations that are fixed during evolution are selectively neutral. This theory was later refined in the nearly neutral theory of evolution which states that also mutations with minor negative effects (nearly neutral) can be fixed by genetic drift.

About at the same time technological advances made it possible to sequence proteins and later nucleic acids: in 1977 the Sanger method to sequence DNA became commercially available. The availability of molecular data for evolutionary studies promoted the development and application of molecular phylogenetic methods. In this field the work of Felsenstein was essential in the development of maximum likelihood methods of phylogenetic inference followed more recently by Bayesian probabilistic methods (Felsenstein 1973, 1981). These are now preferred to parsimony and distance methods because based on probabilistic models of sequence evolution and can be used to test specific evolutionary hypothesis in a solid statistical framework (Lemey et al. 2009).

The second sequencing revolution followed two decades later with the development of Next Generation Sequencing (NGS) methods. NGS boosted the amount of sequence data produced, impacting in many ways how biological research is conducted. Systematic biology, the science that reconstructs evolutionary relationships between species (the Darwinian tree of life) was one of the fields which profited most from this technological advance. NGS facilitated the resolution of complex phylogenetic problems thanks to an increased number of loci that can be analyzed. Furthermore, and most importantly, researcher observed a wildly diffuse incongruence between gene trees. Therefore the habits of interpreting gene trees as species trees became less and less sustainable because there can not be two different species trees given the same set of taxa. As Posada (2016) put it:

"*... the wealth of data resulting form NGS has forced us to stop ignoring phylogenetic incongruence and to reconsider the difference between gene trees and species trees ...*".

Our understanding and awareness of the complexity of the processes that generated the observable genomic diversity (not only the gene diversity, but also non-coding sequences like introns, selfish sequences etc.) has made tremendous progress as both theoretical models and inference methods

advanced under the increasing pressure generated by the unprecedented amount of data. Current methods used in species tree inference incorporate mechanisms such as incomplete lineage sorting, horizontal gene transfer (HGT) and gene duplication and extinction (Maddison 1997). Comparative genomics showed the prominence of these mechanisms and led to the change from the tree of life paradigma to the network or the forest of life (Koonin 2009) (Fig. 3 and 4).

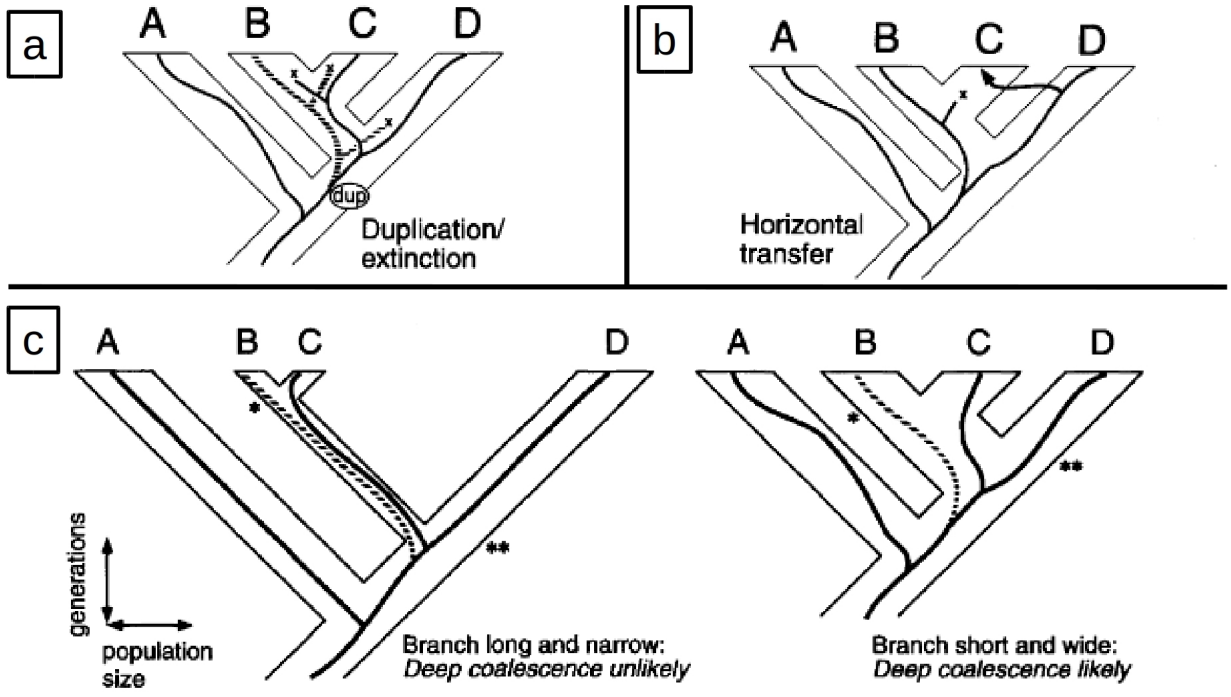


Figure 4. Mechanisms that can cause discrepancies between gene tree and species tree (figure and legend adapted from Maddison 1997) **a)** Gene duplication and extinction (or paralogous sampling). The gene is duplicated to a different locus, indicated by dashed lines. If in descendant species on or the other locus goes extint or is not sampled (x), then the gene tree will disagree with the species tree. **b)** Horizontal transfer. A branch of the gene tree jumps between species lineages. If the indigenous gene copy in the receiving species lineage goes extinct or is not sampled (x), then the gene tree will disagree with the species tree. **c)** Lineage sorting (deep coalescence). Described in a time-forward sense as lineage sorting, an ancestral polymorphism at ** is retained through a lineage to the next speciation event at *, where different forms are sampled in different descendant species. Described in a time-backward sense as deep coalescence, two gene copies from species B and C meet at * but fail to coalesce until deeper than the speciation event at **, at which point the gene from C coalesces first with the gene from D. Failure to coalesce is more likely the shorter (in generation) and wider (in effective population size) the branch is between ** and *.

One major contribution to the analysis of sequence polymorphisms has been the coalescent theory. The coalescent is a population-genetic model fully described by Kingmann (1982) for the first time. The basic idea is that in absence of selection we can model genealogies of sampled gene randomly picking parents in the generation before, when two genealogies pick the same parents they coalesce, eventually all lineages will coalesce in the most recent common ancestor. Mutation can then be added to model randomly "throwing" them on the genealogical tree. The basic model assumes discrete non-overlapping generations, random mating, a large panmictic population, no recombination and no selection. These assumptions probably look unrealistic to most biologists, however the basic model has been extended to include the effects of most of these mechanisms (Fig. 5). Coalescent theory has been extensively used as a mathematical tool for modeling, to do simulation for hypothesis testing and for full-likelihood inference (Rosemberg and Nordborg 2002). Coalescent-based methods are usually used to reconstruct the "neutral" evolutionary history of lineages from sequence data, moreover coalescent simulations provide the null hypothesis to test if patterns of diversity in particular sequences of interest significantly deviate from the null expectation.

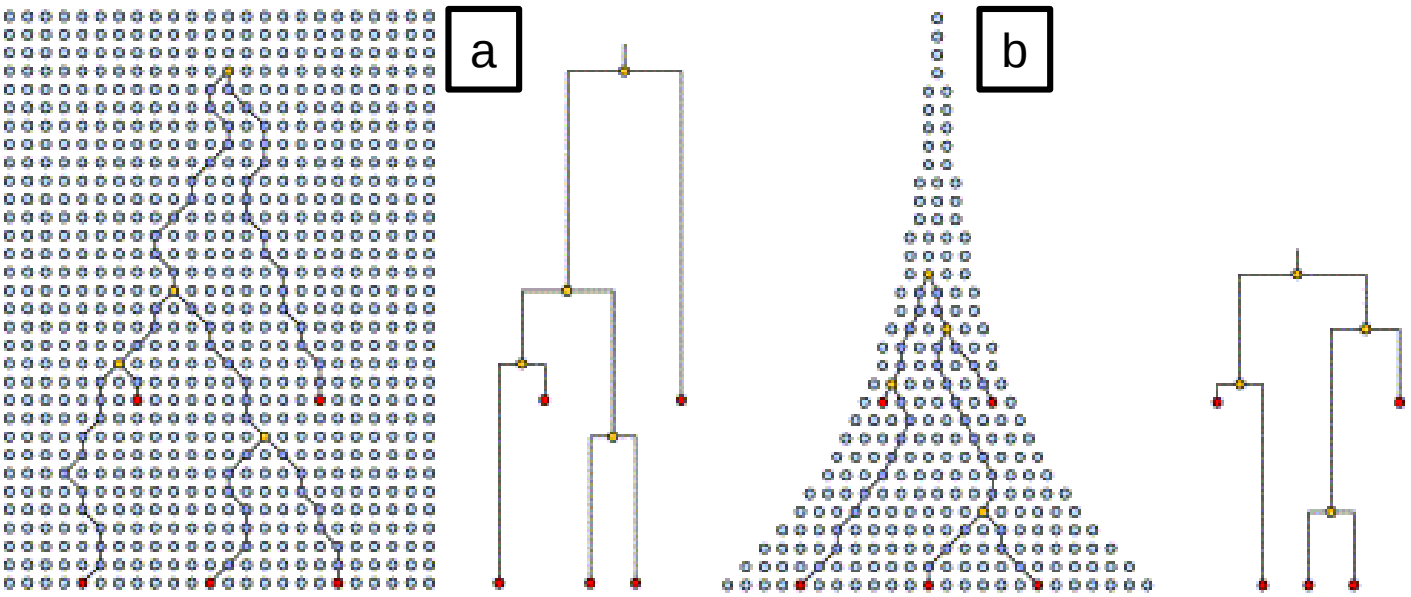


Figure 5. An example of a genealogy of three contemporary genes and two genes sampled 11 generations in the past under two demographic scenarios: **a)** constant population size **b)** population growth (modified from Drummond et al. 2003). The genealogies observable under these two demographic models are different, with genealogies of the population growth model coalescing faster in the most recent common ancestor. Relations between gene genealogies and demography are at the very base of the coalescent theory.

A very short introduction to phytopathology

Plants are normally exposed to thousand of pathogens whose aim is to obtain nutrients at the expenses of the plants. However a single plant species is commonly immune to the vast majority of pathogens, and susceptible to a few. This is the outcome of a million-years arm-race between plant immune systems and pathogens. According to the current paradigm the plant immune systems are mostly based on molecular receptor which perceive pathogen-associated ligands. These receptor can be classified in two groups according to their sub-cellular localization. The first of them is composed by trans-membrane receptors which recognize pathogen-associated molecular patterns (PAMPs) in the apoplast and lead to pathogen triggered immunity (PTI). This level of resistance is normally very effective against most non-adapted pathogens. However, some pathogens can suppress PTI with secreted effector proteins (Jones et Dangl 2006, Giraldo & Valent 2013). Effectors have been suggested to avoid or suppress resistance responses, to promote the spread of the parasite, the nutrient uptake and to give structural support to infection structures (Win et al. 2012, Giraldo & Valent 2013). In response, plants have evolved a second group of defense cytoplasmic receptors, usually with nucleotide-binding and leucine-rich-repeat domains (NB-LRR), which can recognize effectors (then called avirulence factors) and cause effector-triggered immunity (ETI) including defense responses like hypersensitive cell-death (Jones and Dangl 2006). This level of resistance, unlike the PTI is pathogen (and often race) specific, in the sense that a NB-LRR protein specifically recognizes one (or few) avirulence factor(s) present in one or a few pathogen species. Effector-triggered resistance is frequently overcome by new pathogen strains that according to the classical gene-for-gene model, loose or alter the effectors responsible for the recognition (Jones et Dangl 2006, Stergiopoulos and de Wit 2009).

Aim of the thesis

This thesis is one of the outcomes of the NGS revolution in the powdery mildew research. The publication of the wheat and barley powdery mildew genomes in 2010 and 2013 (Spanu et al. 2010 and Wicker et al. 2013) paved the way for large scale comparative and evolutionary analysis of powdery mildew strains. The main aim of my PhD project was to characterize the genome of the newly emerged form of powdery mildew which infects triticale, this work was then expanded to the analysis of genomes of grass powdery mildew strains attacking wild grasses and other cereals and to the attempt to classify and characterize the multitude of effector proteins in *B. graminis.*

CHAPTER 2

Hybridization of powdery mildew strains gives rise to pathogens on novel agricultural crop species

Fabrizio Menardo[1], Coraline R. Praz[1], Stefan Wyder[1], Roi Ben-David[1,2], Salim Bourras[1], Hiromi Matsumae[3], Kaitlin E. McNally[1], Francis Parlange[1], Andrea Riba[4], Stefan Roffler[1], Luisa K. Schaefer[1], Kentaro K. Shimizu[3], Luca Valenti[1], Helen Zbinden[1], Thomas Wicker[1,5] and Beat Keller[1,5]

[1]Institute of Plant Biology, University of Zürich, Zollikerstrasse 107, Zürich 8008, Switzerland

[2] Present address: Institute of Plant Science, ARO-Volcani Center, Bet Dagan 50250, Israel

[3]Institute of Evolutionary Biology and Environmental Studies, University of Zürich, Winterthurerstrasse 190, Zürich 8057, Switzerland

[4]Biozentrum, University of Basel, Klingelbergstrasse 50/70, Basel 4056, Switzerland

[5]Shared last authors

Published in Nature Genetics. Volume 42, Number 2, pages: 201-205. February 2016.

Abstract

Throughout history of agriculture, many new crop species (polyploids or artificial hybrids) were introduced to diversify products or to increase yield. However, little is known of how these new crops impact the evolution of new pathogens and diseases. Triticale is an artificial hybrid of wheat and rye and it was resistant to the fungal pathogen powdery mildew (*Blumeria graminis*) until 2001. We sequenced and compared the genomes of 46 powdery mildew isolates covering several *formae speciales*. We found that *B.g. triticale* growing on triticale and wheat is a hybrid between wheat powdery mildew (*B.g. tritici*) and mildew specialized on rye (*B.g. secalis*). Our data show that the hybrid of the two mildews specialized on two different hosts can infect the hybrid plant species originating from those two hosts. We conclude that hybridization between mildews specialized on different species is a major mechanism of adaptation to new crops introduced by agriculture.

Results and Discussion

The artificial hybrid triticale was introduced into commercial agriculture in the 1960's. The hexaploid triticale genome is composed of genomes A and B from wheat plus the rye genome R (AABBRR)(Oettler 2005). Triticale was initially resistant to powdery mildew, however in 2001 the disease was first observed on this crop and it has since become a major disease in Europe (Mascher et al. 2005, Walker et al. 2011,). We sequenced 46 isolates of *B. graminis* (including the 96224 reference isolate, Wicker et al. 2013) from different European countries and Israel (Supplementary Table 1) with host ranges corresponding to four different *formae speciales* (*ff.spp.*), including the previously described rye (*B.g. secalis*) and wheat (*B.g. tritici*) powdery mildews. Our infection tests showed that *B.g. secalis* grows exclusively on rye while *B.g. tritici* is able to grow on tetraploid (durum) and hexaploid (bread) wheat (Table 1, Supplementary Note A, Supplementary Table 2, Supplementary Figs. 1-23). Based on infection tests, we define the *f.sp. triticale* as able to grow on triticale, hexaploid and tetraploid wheat (with lower penetration efficiency, Supplementary Note B, Supplementary Fig. 24), and to a very limited extent on rye. Furthermore, we designate mildew growing exclusively on tetraploid wheat as new *f.sp. B.g. dicocci* (Eshed et al. 1994).

Table 1. Host specificity of powdery mildew *ff.spp.* Columns represent different host species, rows different *ff.spp.* of *Blumeria graminis*. A compatible interaction is indicated by (+), an incompatible interaction by (-). Intermediate phenotypes are indicated by (+/-), referring to *B.g. triticale* isolates that show limited growth on rye. [-/(+)] indicates a very reduced growth that we observed for some *B.g. secalis* isolates on triticale.

	Tetraploid wheat	Hexaploid wheat	rye	triticale
B.g. dicocci	+	-	-	-
B.g. tritici	+	+	-	-
B.g. secalis	-	-	+	-/(+)
B.g. triticale	+	+	-/+	+

Overall, the genomes of the 46 isolates are very similar to each other, allowing a high-quality mapping of the 45 re-sequenced isolates to the reference isolate 96224 (Supplementary Note C). We identified between 115,543 and 332,450 polymorphic nucleotide sites per isolate when compared to the *B.g. tritici* reference genome (Supplementary Note D, Supplementary Figs. 25-26). A principal component analysis based on 717,701 polymorphic sites clearly distinguished four groups (Fig. 1) which correspond to the four *ff.spp.* identified in the infection tests. This indicates that gene flow is restricted between *ff.spp.* Interestingly, the *B.g. triticale* isolates formed a group distinct from the *B.g. tritici* isolates. This contradicts the current hypothesis of a host range expansion of *B.g. tritici* to triticale through mutation of a few genes (Walker et al. 2011, Troch et al. 2012 and 2013). Instead, our analysis shows that *B.g. triticale* isolates form a specific group with a distinct evolutionary history (Supplementary Note E, Supplementary Figs. 27-30).

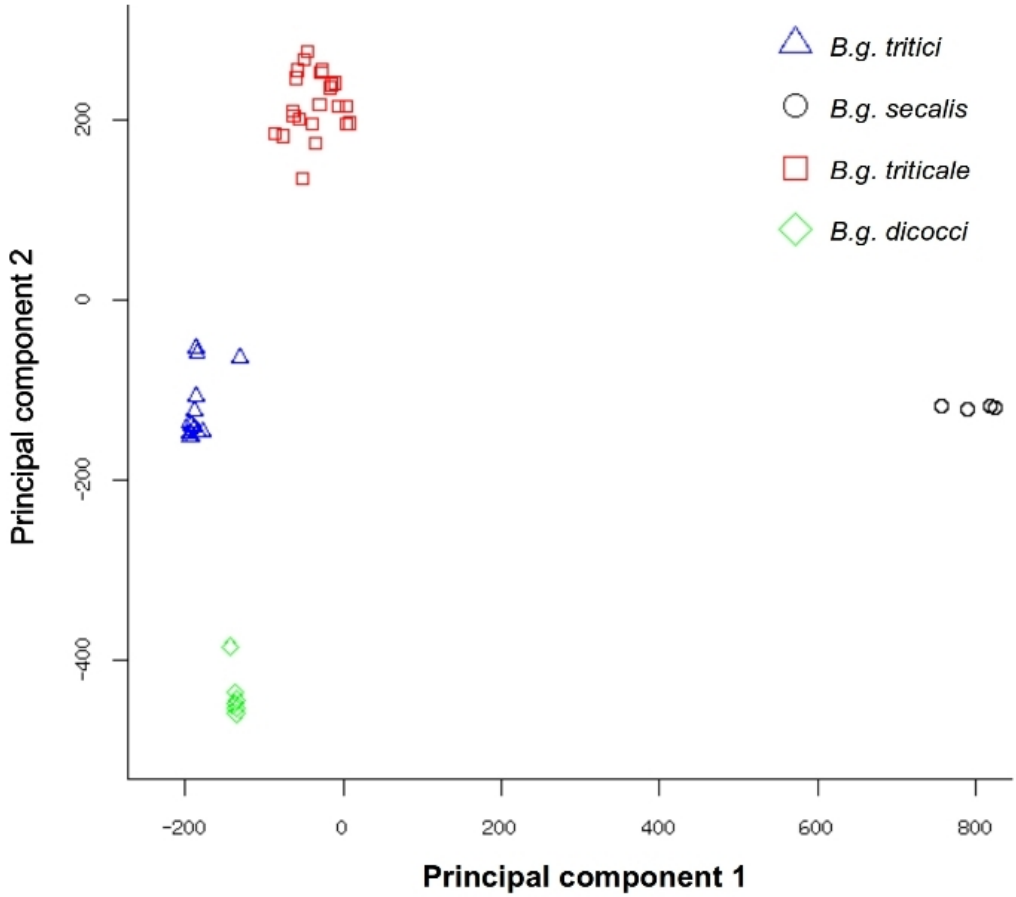


Fig. 1. Analysis of sequence diversity in genomes of 46 sequenced *B. graminis* isolates. Principal component analysis (PCA) based on 717,701 SNPs that are polymorphic in at least two isolates. The PCA differentiates four groups corresponding to the four different *ff.spp.*, which are defined by their differential host ranges.

For genome comparisons, we identified sets of polymorphisms that are fixed in the four *ff.ssp.* (i.e. polymorphisms shared among all isolates of a *f.sp.*). Such fixed polymorphisms between *ff.spp.* (i.e. substitutions) will hereafter be referred to as the "genotype" of a *f.sp*. Interestingly, the *B.g. triticale* genome consists of large genomic segments that have the same genotype as *B.g. secalis*, alternating with segments with a *B.g. tritici* genotype (Figs. 2a and 2b, Supplementary Figs. 31-33). Such patterns are characteristic for recent hybrids, reflecting a limited number of recombination events between parental genotypes. Furthermore the analysis of nucleotide diversity in the genome of the different *ff.spp.* showed that *B.g. triticale* has a characteristic distribution with multiple peaks that is consistent with the hypothesis of a recent origin by hybridization (Figs. 2c and 2d, Supplementary Notes F and G).

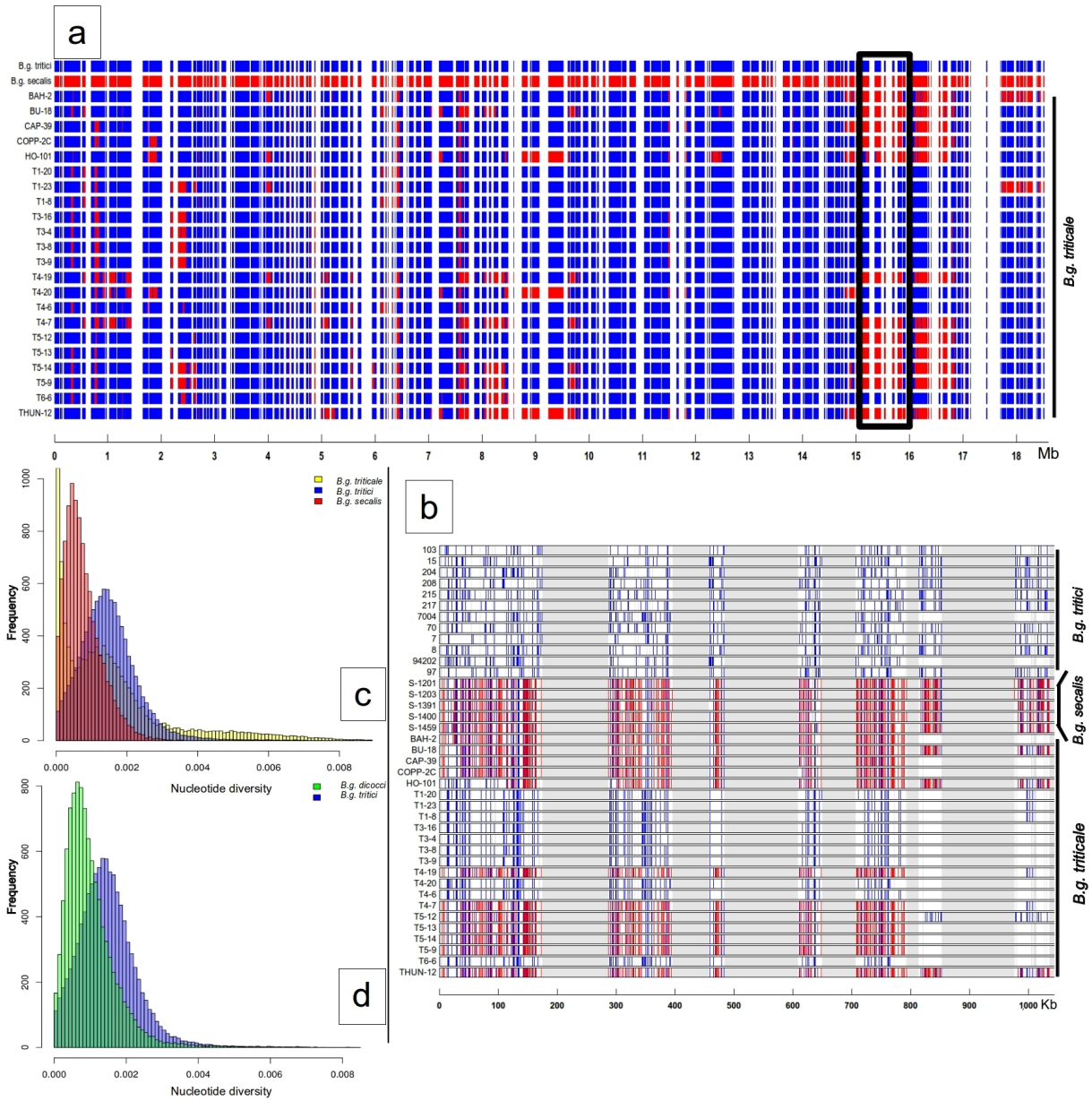


Fig. 2. Examples of nucleotide diversity patterns in powdery mildew isolates. a) Nucleotide substitutions in *B.g. triticale* isolates compared to *B.g. tritici* and *B.g. secalis* on the *B. graminis* linkage group 3 (Bourras et al. 2015). Fixed polymorphisms specific for *B.g. tritici* isolates are depicted in blue while those specific for *B.g. secalis* isolates are depicted in red. The black box shows the section of the linkage group represented in Fig. 2b. **b)** Polymorphism pattern on physical contig 140 on linkage group 3 in *B.g. tritici* (top 12 lines), *B.g. secalis* (following 5 lines) and *B.g triticale* isolates (last 22 lines). Polymorphisms compared to the *B.g. tritici* reference genome sequence 96224 that are present in at least one of the *B.g. tritici* isolates are colored in blue in all isolates of all *ff.spp.* Polymorphisms not present in *B.g. tritici* but in *B.g. secalis* are colored in red in *B.g. secalis* and *B.g. triticale* isolates. Grey windows represent sequence gaps in the *B.g. tritici* reference genome, white areas represent non-polymorphic regions. **c)** Histograms of nucleotide diversity (π) of aligned genome windows larger than 10 kb between *B.g. triticale* isolates (yellow), *B.g. tritici* isolates (blue) and *B.g. secalis* isolates (red) (Supplementary Note G). **d)** Histograms of nucleotide diversity (π) of aligned genome windows (larger than 10 kb) in *B.g. tritici* isolates (blue) and *B.g. dicocci* isolates (green).

The distribution of *B.g. tritici* is shifted to the right compared to the one of *B.g. dicocci*, indicating a greater diversity in *B.g. tritici* than in *B.g. dicocci*.

To test this hypothesis, we used 172,274 fixed polymorphisms that distinguish *B.g. tritici* from *B.g. secalis*. We found that in *B.g. triticale* between 11.9 % and 21.4 % of the polymorphic sites represent the *B.g. secalis* genotype. These polymorphic sites were located on genomic segments that make between 6.6% and 17.3% of the genomes of the *B.g. triticale* isolates. In contrast, over 80% of their genomes are of the *B.g. tritici* genotype. Thus, we conclude that *B.g. triticale* is a hybrid of *B.g. tritici* and a *B.g. secalis* and that this hybridization happened very recently (Supplementary Note H, Supplementary Table 3).

Furthermore, phylogeographic analysis and the proportions of parental genomes suggest that the initial hybridization event was followed by 2 back-crosses with *B.g. tritici* and that this process likely occurred in Europe (Supplementary Figs 34-37 Supplementary Note I).

These findings raised the question whether *B.g. triticale* originated from a single hybridization of *B.g. secalis* and *B.g. tritici* (followed by two back-crosses), or more individuals of the two *ff.spp.* contributed. If the first hybridization was a single event that involved only one *B.g. s*ecalis isolate we expect to observe no diversity in the *B.g. triticale* portion of the genome inherited from *B.g. secalis*. Based on phylogenetic analysis we found 66 genes for which all *B.g. triticale* isolates inherited the *B.g. secalis* gene. We counted the number of haplotypes present in *B.g. triticale* for each of these 66 genes and found six genes that show two different haplotypes. These different haplotypes are also present in the 5 sequenced *B.g. secalis* isolates. We conclude that the minimum number of *B.g. secalis* individuals that contributed to *B.g. triticale* is two, and this is likely due to two independent origins of *B.g. triticale* (Supplementary Note J, Supplementary Fig.38).

Blumeria graminis, as most fungi, has two mating types (Wicker et al. 2013). Since mating types of *B.g. secalis* and *B.g. tritici* have different genotypes we traced back the mating types origin in *B.g. triticale*. We found that all *B.g. tritici* partners in the first hybridizations seem to have been of the MAT1-1-3 mating type while *B.g. secalis* partners were of the MAT1-2-1 mating type. The MAT1-2-1 locus of the *B.g. tritici* genotype was acquired by *B.g. triticale* through one of the back crosses, while the MAT1-1-3 locus of the *B.g. secalis* genotype is not present in any of the sequenced *B.g. triticale* isolates (Supplementary Note K, Supplementary Figs. 39-40). These findings are in contrast to those on the plant pathogen *Zymoseptoria pseudotritici* whose origin was traced back to a single hybridization event (Stukenbrock et al. 2012). Furthermore, unlike *Z. pseudotritici*, the new *f.sp. B.g. triticale* did not likely pass through a bottleneck, because the multiple isolates involved in the hybridizations passed on a considerable part of parental genetic diversity (Supplementary Note F, Supplementary Table 4).

Based on the number of observed recombination between *B.g. tritici* and *B.g. secalis* genotypes we estimated that *B.g. triticale* isolates underwent between 7 and 47 (depending on the isolates) sexual cycles from the origin of the *f.sp.* (Supplementary Notes L and M, Supplementary Fig. 41). Since *B. graminis* has a maximum of one sexual cycle per year (Wicker et al. 2013) we conclude that *B.g. triticale* originated after introduction of triticale as a commercial crop in the 60's, possibly multiple times and since then different isolates underwent a different number of sexual generations. Since *B.g. triticale* is a very recent hybrid we studied the phylogenetic relationship of its two parents. We estimated the divergence time between the *B.g. triticale*'s parental *ff.spp. B.g. secalis* and *B.g. tritici* using 206 orthologous single copy genes and found that they diverged between 168,245 and 240,169 years ago (Supplementary Note N, Supplementary Fig. 42, Supplementary Table 5). In contrast, their hosts rye and wheat diverged already approximately 4 million years ago (Middleton et al. 2014). This incongruence between divergence of host and pathogen can be due to host tracking: until a few hundreds of thousands of years ago wheat and rye were still close enough to be in the host range of a single *f.sp.* and the divergence in to two distinct *ff.spp.* occurred only relatively recently. This would be consistent with previous findings that the *ff.spp. B.g. tritici* and *B.g. hordei* diverged approximately 6 million years ago (Wicker et al. 2013), while their hosts barley and wheat diverged at least 2 million years earlier (Middleton et al. 2014). An alternative explanation could be a recent host jump from wheat to rye, or vice versa, followed by rapid emergence of barriers to gene flow.

Despite our massive sequencing effort, the molecular basis for host range expansion of *B.g. triticale* remains obscure. Genetic determinants for the expanded host specificity to triticale must be located on genomic segments that were inherited from *B.g. secalis*. Among these genes there might be the genetic determinant for the host range expansion of *B.g. triticale*. Six of the 66 *B.g. secalis* genes inherited by all *B.g. triticale* isolates genes encode putative effectors (i.e. proteins which are secreted into the host cell to facilitate pathogen proliferation) (Supplementary Note O, Supplementary Table 6). In *B. graminis* they are typically small proteins with short characteristic motifs identified with bioinformatics criteria (Wicker et al. 2013) (i.e. presence of signal peptide and lack of homology with protein domains of other organisms). However, transcriptome profiling of *B.g. triticale* on wheat and triticale showed that none of these genes were differentially expressed by the fungus on the two different hosts. In general, the transcriptome profile of *B.g. triticale* was mostly independent of the host species (Supplementary Note P, Supplementary Figs. 43-47). Alternatively, host specificity could be a quantitative trait that requires a certain number of genes with partially overlapping and/or complementary functions. Effector genes in particular are thought to have overlapping or partially redundant functions (Birch et al. 2008).

It is intriguing that *B.g. triticale's* host (triticale) is a hybrid of rye and bread wheat, which are inversely the hosts of *B.g. triticale*'s parents. Interestingly, bread wheat (*B.g. tritici*'s host) itself is the result of a hybridization approximately 10,000 years ago between domesticated tetraploid (emmer) wheat and the diploid wild grass *Aegilops tauschii* (Salamin et al. 2002). Moreover the host range of *B.g. tritici* (hexaploid and tetraploid wheat) includes the host range of *B.g. dicocci* (tetraploid wheat). This is reminiscent of the host range expansion of *B.g. triticale* that includes the host range of one of the parent species *B.g. tritici* (wheat), in addition to triticale. We therefore hypothesized that *B.g. tritici* could also be a hybrid between a pathogen of tetraploid wheat (i.e. *B.g. dicocci*) and one that infects *Aegilops tauschii.* However, in contrast to *B.g. triticale*, we did not find large genomic segments in *B.g. tritici* isolates that could be assigned to a *B.g. dicocci* genotype. This could be explained by multiple rounds of recombination that have eroded the characteristic pattern of sequence segments. A different approach to test hybridization that is more resistant to the action of recombination makes use of gene genealogies (Hein et al. 2005) which have been used to identify hybridization in different organisms (Xu et al. 2000, Sota 2002). We used coalescent-based methods to calculate the probability of the evolutionary model in which both *B.g. tritici* and *B.g. triticale* are hybrids, and of four alternative models in which there is no hybridization or only one of the two *f.sp.* is a hybrid (Degnan et al. 2005, Than et al. 2008, Yu et al. 2011 and 2012). The model depicted in Fig. 3 with two hybridizations resulted to be the most likely (Supplementary Note Q, Supplementary Tables 7-8, Supplementary Fig. 48). However, we cannot completely rule out that *B.g. tritici* originated through a more complex, not tested, scenario with multiple past admixture events.

We conclude that *B.g. tritici* arose sometime after the formation of bread wheat, probably several thousand years ago, from the hybridization between a *B.g. dicocci* strain and a different, yet unknown mildew strain. Hybridization has been reported as important for adaptation to new hosts in several fungal and oomycete pathogens (Brasier et al. 1999, Brasier 2000, Newcombe 2000,Brasier et al. 2010, Goss et al. 2011, Farrer et al. 2011). Our data now show that hybridization was the causal step for host range expansion to a newly bred or evolved species. It is apparent that co-evolution based on hybridization is a likely evolutionary pathway for *B. graminis* that infects wheat and other grasses, which themselves evolve predominantly through hybridization (Feldman and Levy 2012). It is particularly fascinating that pathogen evolution mirrors the evolution on the host side, and the hybrid of two mildews specialized on two different hosts can infect the hybrid plant species originating from those two hosts (Fig. 3). It is possible that our findings define a more general evolutionary pattern: stem rust *Puccinia graminis* was reported to infect triticale approximately at the same time as *B. graminis* (Tian et al. 2004).

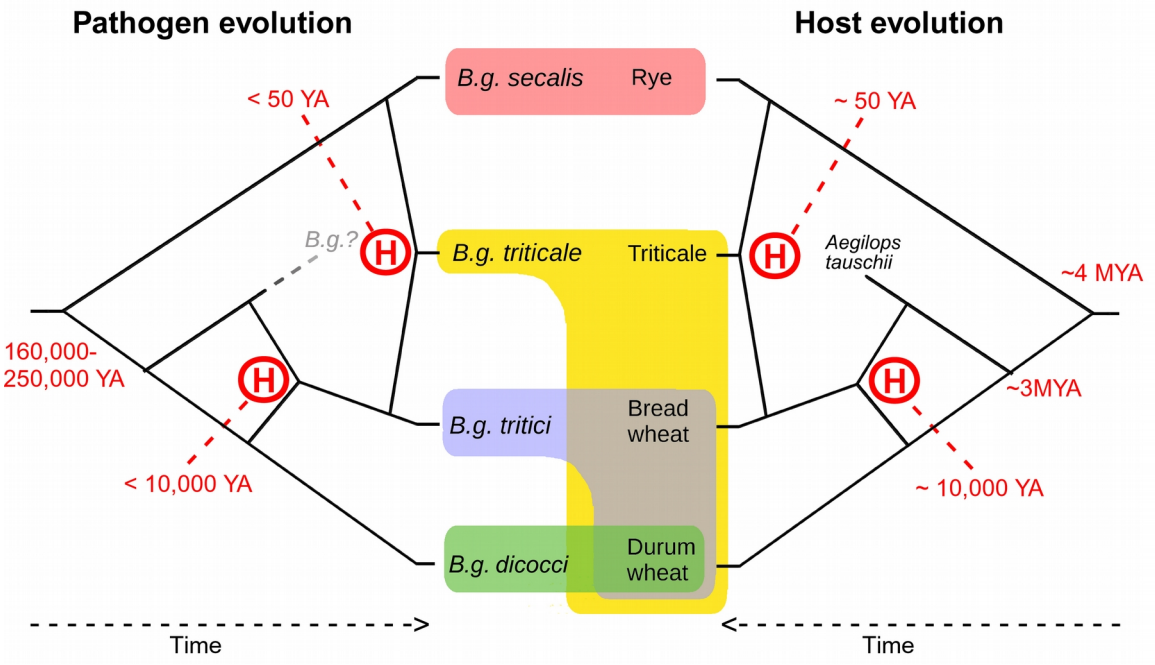


Fig. 3. Model for the evolution of specialized forms and host ranges in *Blumeria graminis*. Phylogenetic trees of *Blumeria graminis* and its hosts are represented facing each other, revealing that pathogen evolution mirrors the evolution of the host species. The branch corresponding to the unknown *f.sp.* (*B.g. ?*) that hybridized with *B.g. dicocci* to form *B.g. tritici* is shown in gray. Hybridization events are marked with H. Estimated times of species and *ff.spp.* divergences and hybridizations events are shown in red (YA: years ago, MYA: million years ago). Origin times of *B.g. triticale* and *B.g. tritici* are based on the origin of the host species. The host ranges of the different *ff.spp.* are indicated as shaded areas of different colors.

Also in *P. graminis* hybridization between *ff.spp.* has been reported to be important for host range determinatn (Luig and Watson 1972). It is noteworthy that this pattern of evolution is observed in pathogens of agricultural species. Possibly, agricultural ecosystems increase the possibility that different pathogens co-occur in large populations, thus making hybridization more likely than in natural ecosystems. Many agricultural crops are polyploids and/or contain resistance genes introgressed from wild relatives. Little is known what determines the durability of such introgressed genes, but the hybridization of pathogens of the wild relative with the crop pathogen should be considered in resistance breeding approaches. In addition, in the future production of man-made

crops, genetic resistance of parents should be carefully selected to make the resulting hybrids more durably resistant in the field.

Methods

Host specificity tests

To test the host specificity of the isolates used in this work we infected six cultivars of triticale, five cultivars of wheat (three hexaploid and two tetraploid) and three cultivars of rye with 46 *Blumeria* isolates (Supplementary Table 2). The plant cultivars used in the host specificity tests were chosen for the absence of known race-specific resistance genes and their known high susceptibility to many tested mildew isolates. In this way we avoided race-specific resistance as a confounding factor in determination of host specificity. We infected 10 day-old detached leaf segments with fresh spores and the infected leaf segments were kept on benzimidazole agar plates at 20 °C, 70% humidity and in 16h light / 8h dark conditions. We scored the phenotypes ten days after infection using three categories: virulent, intermediate and avirulent (Supplementary Fig. 1-23, Supplementary Note A).

Staining protocol for assessment of penetration efficiency

Infected leaf segments were collected at 2 days post infection, and incubated for 4 days in destaining solution (8.3% lactic acid, 16.7% glycerol in ethanol). Leaves were stained for 45 seconds with neutral red (0.1%), this was used as a contrasting agent to observe haustoria inside plant epidermal cells. Aerial fungal structures were stained for 45 seconds using 0.25% coomassie blue.

Sequencing, mapping, SNP call, assemblies and principal component analysis (PCA)

DNA extraction was performed following the methods of Bourras and colleagues (2015). 101 bp paired-end libraries were created and sequenced with Illumina Hi-Seq at the Functional Genomic Center of Zürich, or at The Genome Analysis Centre Norwich Research Park (TGAC), Norwich, UK obtaining between 31,039,416 and 62,834,815 reads (mean 41,810,764) (Supplementary Table 1). Reads were mapped on the *B.g. tritici* reference genome[4] using Bowtie 2.1.0 (Langmead and Salzberg 2012) with option --score-min L,-0.6,-0.25. We used the following command of samtools 0.1.19 (Li et al. 2009) to convert formats and collect information about single genomic positions: view, sort, mpileup -q 15, (only reads with mapping quality greater than 15 are considered). Finally we used bcftools to generate a vcf file that was parsed with home-made perl scripts (available upon request). We considered as high confidence SNPs only positions with minimum mapping score of

20, minimum coverage of 20 and minimum frequency of the alternative call of 0.95. Polymorphisms were considered as "fixed" if they were found to be present in all isolates of a *f.sp.* and different in the other *ff.spp.*

De novo assemblies of all isolate genomes were performed with CLC Genomic Workbench 7.5 with standard parameters. PCA was performed with the R package GAPIT (Lipka et al. 2012).

Identification of orthologous genes

We used gmap (version 2013-07-20) (Wu and Nacu 2010) to annotate genes in the assembly of isolate S-1201 (as *B.g. secalis* reference genome) using the 7,186 genes of the *B.g. tritici* reference genome (isolate 96224) as a template. We found 6,864 genes in the genome assembly of isolate S-1201. The protein and CDS databases of *B.g. hordei* and *N. crassa* were downloaded from blugen.org and broadinstitute.org (accessed 01/03/2014). After elimination of genes with homology to the *Blumeria* repeat database (BRD) and to the *Triticeae* transposable elements (PTREP12) databases we retained 9,733 genes for *N. crassa* and 6,011 for *B.g. hordei.* To cluster genes in families we used all against all blast search (Altschul et al. 1997)) and we grouped together genes that hit each other with a minimum alignment length of 150 bp and a blastn e-value $\leq 10^{-10}$. We then defined as single-copy gene families all families with four genes, one for each of the species used (*N. crassa, B.g. tritici, B.g. hordei, B.g. secalis*), that hit reciprocally each other as first blast hit. With these criteria we obtained 208 orthologous gene families. We used gmap (version 2013-07-20) to retrieve these genes in all the *B.g. tritici* and *secalis* isolates, using as template the genes from the same *f.sp.* (isolates 96224 and S-1201). Assemblies for the *B.g. hordei* isolates AOIY01 and AOLT01 were obtained from Hacquard and colleagues (20113) We annotated the 208 genes on these two assemblies with gmap (version 2013-07-20) using the genes of the *B.g. hordei* reference isolate as template. We could not find two genes in all *B.g. hordei* isolates, therefore we excluded them. The final dataset was composed of 206 orthologous single-copy genes from *N. crassa*, 3 *B.g. hordei* isolates, 5 *B.g. secalis* isolates and 13 *B.g. tritici* isolates. This dataset was used to infer species tree phylogeny and estimate divergence time (Supplementary Note N).

We used analogous methods to define single copy orthologous genes in *B. graminis* (without *N. crassa*). We used all against all blast search (between *B.g. hordei* (reference isolate), *B.g. tritici* (96224), *B.g. secalis* (S-1201) and *B.g. dicocci* (220)). We grouped together genes that hit each other with a minimum alignment length of 150 bp and a blastn e-value $\leq 10^{-10}$. We then defined as single-copy gene families all families with four genes, one for each of the species used (*B.g. dicocci*, *B.g. tritici*, *B.g. hordei*, *B.g. secalis*), that hit reciprocally each other as first blast hit. This resulted in 4,556 single copy genes. We then retrieved these genes in all isolates, using as template

the genes from the same *f.sp.* (isolates 96224 for *B.g. tritici*, S-1201 for *B.g. secalis*, 220 for *B.g. dicocci*), the reference isolate of *B.g. hordei* was used as outgroup. Multiple alignments were generated with muscle (Edgar 2004) and we inferred maximum likelihood trees for all alignments with RAXML 8.0.22 (Stamatakis 2014) using a GTR + GAMMA model Tavarè 1986, Yang 1993 and 1994). We used Newick Tools (Junier and Zdobnov 2010) to identify particular topology patterns. This dataset was used in phylogeography analysis (Supplementary Note I), in the identification of genes inherited from *B.g. secalis* in all *B.g. triticale* isolates (Supplementary Note O), in the estimation of the minimum number of isolates that contributed to the first hybridization (Supplementary Note J) and in the coalescent based estimation of the most likely evolutionary network (Supplementary Note Q).

Alignments and phylogeny for divergence time estimation

Protein multiple alignments were performed with Muscle 3.8.31 and retrotranslated into nucleotides with TranslatorX v1.1 (Abascal et al. 2010). The concatenation of all the alignments was 410,189 bp long. Phylogenetic inference of the partitioned dataset was performed with MrBayes 3.2.2 (Romquist et al. 2012). We ran two independent replications of 10,000,000 generations. Variation of substitution rates across sites was modeled with a discretized (4 categories) gamma (Γ) distribution. The chains have been let free to sample all models of the GTR model family using reversible jump Monte Carlo Markov Chain (Huelsenbeck et al. 2004). The node that links *B.g. hordei* and *B.g. tritici* was used as calibration point (5.2 - 7.4 million years ago (Wicker et al. 2013) or 10,000-14,000 years ago (Wyand and Brown 2003) (Supplementary Note N) under the independent gamma rate relaxed clock model (white noise model) (Lepage et al. 2007).

RNA extraction and transcriptome analysis

Leaf segments of the wheat variety Chinese Spring and the triticale variety Timbo were infected with two different *B.g. triticale* isolates (THUN-12 and T3-8). Leaves were left on Benzimidazole agar plates as described in Parlange et al. (2011) and harvested 2 days post infection. Each pathogen / host combination was replicated three times. RNA extraction was performed with the miRNeasy mini kit from Qiagen according to the manufacturer recommendations. 125 bp single-end libraries were created and sequenced with Illumina Hi-Seq at the Functional Genomics Center Zürich. RNAseq reads were mapped with STAR (Dobin et al. 2013) (allowing four mismatches for 100 bp). Read counts were determined with featureCounts 1.4.6 (Liao et al. 2014). The R package edgeR was used for statistical analysis and genes were tested for differential expression with a generalized linear model and tagwise estimation of dispersion (Robinson et al. 2010).

Test of evolutionary networks with PhyloNet

Probabilities of five different evolutionary hypotheses (networks a, b, c, d and e, graphically represented in Supplementary Fig. 35) were evaluated with a coalescent-based method implemented in PhyloNet (Than et al. 2008) (Supplementary Note Q). Because of computational limitations we produced two subsets of isolates and gene trees. We tested two different, non-overlapping sets of isolates (one *B.g. hordei*, two *B.g. secalis*, three *B.g. tritici*, *dicocci* and *triticale*, for a total of 12 isolates for each dataset), each of them with 10 different sets of 300 randomly selected single copy gene trees inferred with RAXML. The composition of the first dataset is: *B.g. hordei* reference isolate; 96224, 97 and 103 (*B.g. tritici*); S-1201 and S-1400 (*B.g. secalis*); THUN-12, T3-8 and COPP-2C (*B.g. triticale*); 58, 66 and 220 (*B.g. dicocci*). The composition of the second dataset is: *B.g. hordei* reference isolate; 94202, 8 and 70 (*B.g. tritici*); S-1203 and S-1459 (*B.g. secalis*); T1-23, T6-6 and HO-101 (*B.g. triticale*); 63, 207 and 209 (*B.g. dicocci*). To compare the likelihood of models with different number of parameters we used three information criteria, Akaike information criterion (AIC) (Akaike 1974), the corrected AIC (AICc) and the Bayesian information criterion (BIC) (Schwarz 1978). These measures give increasing penalties to the models containing more parameters.

Estimation of the number of sexual generation from hybridization in *B.g. triticale*

We estimated the number of sexual generations after hybridization for each *B.g. triticale* isolate using the following formula modified from Stuckenbrok et al. (2012) (the individual components are explained below):

Sexual generations (SG) = ((NR*BGSS) / (2*BGSG)) *0.22 (1.1)

NR : number of recombination break points
BGSS : Average length of B.g. secalis segments after two back-crosses (3.3Mbp)
BGSG : amount of B.g. secalis genome in the B.g. triticale isolate in bp.

The original formulation in Stukenbrock et al. (2012) is :

SG = (ALRWFH)/(ALORW)

(1.2)

ALRWFH : Average length of recombinant windows after first hybridization
ALORW :Average length of observed recombinant window

We used equation 1.2 for the calculation of BGSS. We estimated the recombination rate from the genetic consensus map of *B.g. tritici* which has a length of approximately 1,800 cM (Bourras et al. 2015). Thus, one expects one recombination event every 10 Mbp in a sexual cycle. Therefore, after the first hybridization, we expect segments of *B.g. secalis* genome sequences with an average length of 10 Mbp. In the first back cross, they recombine on average once and the resulting in parental segments of 5 Mbp. With the second back-cross on average each window recombines 0.5 times. This results in BGSS, the average length of 3,333,333 bp for *B.g. secalis* segments in *B.g. triticale* after two back-crosses.

For each further generation *B.g. triticale* isolates mate between each other, the probability that a new recombination event can be observed is given the probability that recombination occurs in a region that is inherited from *B.g. secalis* in one isolate and *B.g. tritici* in the other. Assuming a proportion of 12.5% of *B.g. secalis* genome in *B.g. triticale* isolates this is equal to 0.22 (i.e. 0.125*0.875*2).

Adding these terms equation 1.2 results in:

SG = (BGSS/(Average length of *B.g. secalis* windows))*0.22. (1.3)

Since the *B.g. tritici* reference genome sequence is fragmented (due to its extremely high repeat content) (Supplementary Note G) we cannot use the average length of *B.g. secalis* windows (most of them would be delimited by the end of the contig). However detection of recombination breakpoints (NR) is not affected by contig size. The relation between average length of *B.g. secalis* windows and NR is given by :

Average length of *B.g. secalis* windows = (2*BGSG/NR) (1.4)

If we substitute equation 1.4 in equation 1.3 we obtain equation 1.1.

We defined putative recombination break points as borders between regions that contain fixed substitution inherited from the *B.g. secalis* parent and such inherited from the *B.g. tritici* parent. Here, we excluded stretches of fixed polymorphisms that are likely the result of gene conversion events. Gene conversions typically affect fragments of a less than 1,000 bp. Thus, we ignored groups of polymorphisms of a different parental genotype if they were clustered in a region of less than 1,000 bp in size, or single substitutions of one genotype in a region otherwise originating from the other genotype.

The proportion of the genome that was contributed by *B.g. secalis* (BGSG) was calculated by adding up the sizes of sequence contigs which only contain polymorphisms of the *B.g. secalis* genotype (Supplementary Note F).

Accession codes

All genomic and transcriptomic sequences used for this study can be accessed at the Sequence Reads Archive or at GEO with accession number SRP062198 and GSE73399.

Acknowledgments

We want to thank Susanne Brunner for advice and support, Prof. Dinoor, Veronique Troch and Fabio Mascher for mildew isolates, Catharine Aquino and Hubert Rehrauer for their help with sequencing and bioinformatic analysis, Alex Widmer for reading and commenting the manuscript. This work was supported by the University Research Priority Programme (URPP) 'Evolution in Action' of the University of Zurich, the Swiss National Science Foundation grant 310030B_144081/1 and an Advanced Investigator Grant from the European Research Council (ERC-2009-AdG 249996, Durable resistance).

Author contributions

TW and BK designed the project and wrote the manuscript, TW, SR and FM wrote software for the analysis, AR and FM performed statistical analysis, RBD, CP, LV, KM and FM phenotyped and propagated the isolates, RBD and FM collected isolates and extracted DNA for sequencing, FP and FM performed crosses between mildew isolates, HZ performed RNA extraction, KKS discussed the population genetic analysis, LS performed staining of infected leaves, SAM developed the staining protocol, CP and FM analyzed RNAseq data, SW, HM, TW, SR, CP and FM performed bioinformatics and population genetic analysis.

Conflict of interests

The authors declare that there are no conflicts of interests.

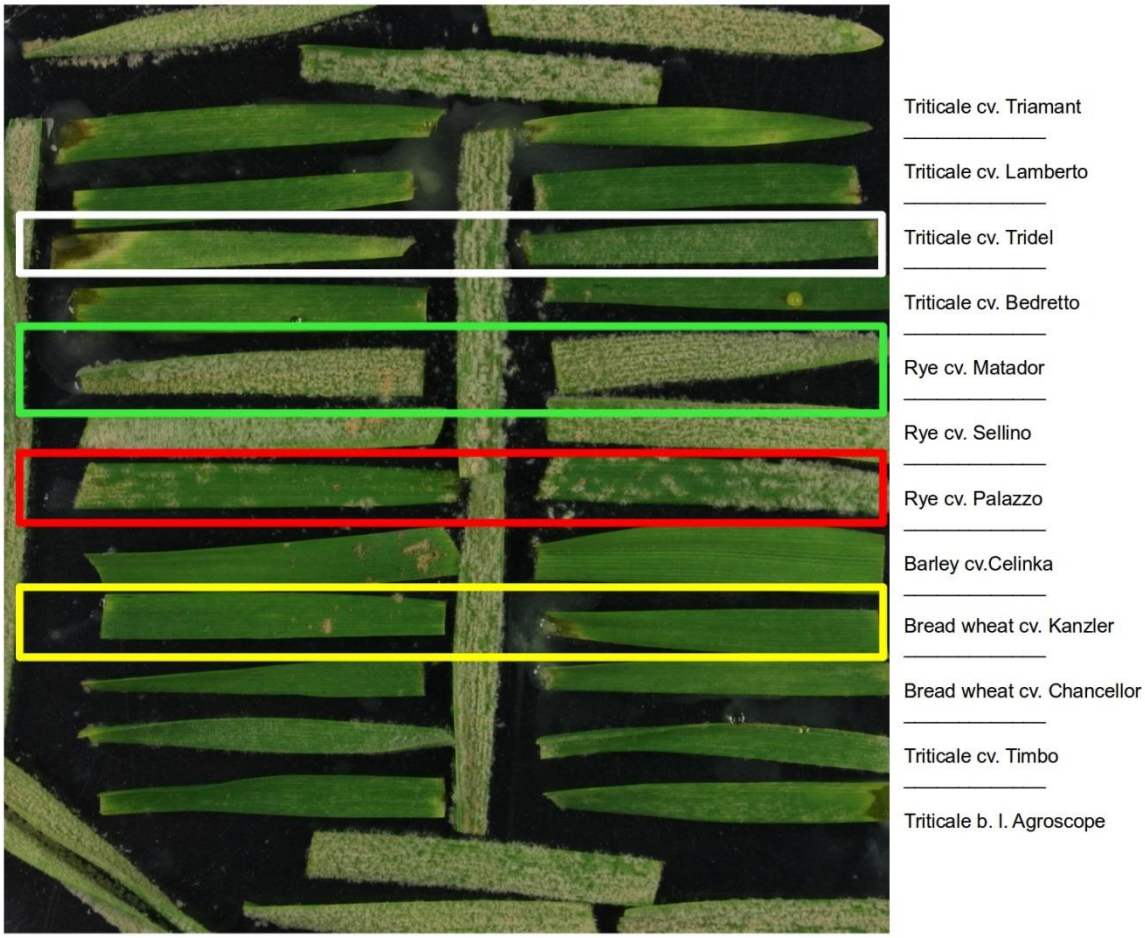
URLs

Triticeae transposable elements (PTREP12) databases:
http://botinst.uzh.ch/research/genetics/thomasWicker/TREP

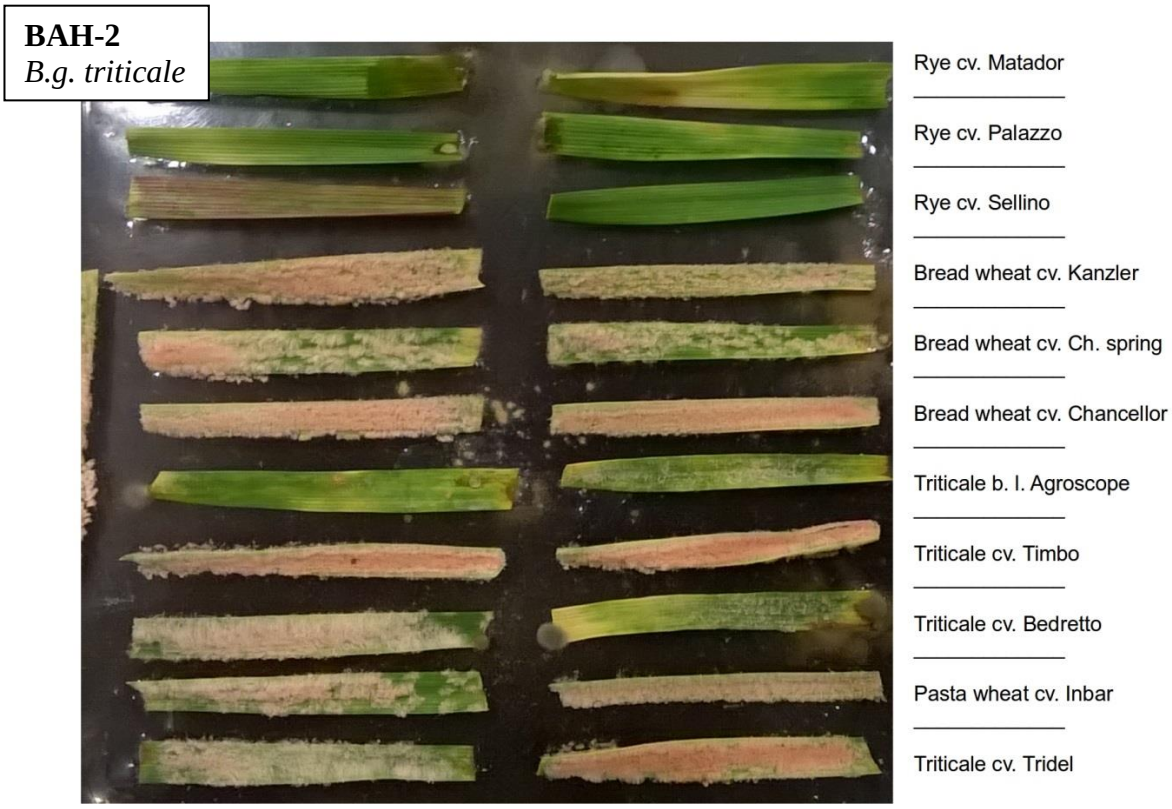
Supplementary Information

Content

I Supplementary Figures



Supplementary Fig. 1. Example for phenotyping of *Blumeria graminis* isolates on seedling leaf segments from different cereal species. The photograph shows a plate infected with the *B.g. secalis* isolate S-1203. The two columns are two biological replicates of the same set of cultivars (cv.), and are surrounded by the susceptible rye cultivar Matador as a positive control for even infection. The name of the plant cultivar is reported on the right, Agroscope is a breeding line obtained from Agroscope Nyon, Switzerland. The white rectangle highlights an example of a very reduced growth on the triticale cv. Tridel. Such phenotypes were considered avirulent. The green rectangle highlights the virulent phenotype on the rye cv. Matador. The red rectangle highlights an example of an intermediate phenotype on the rye cv. Palazzo. The yellow rectangle highlights an example of an avirulent phenotype on the wheat cv. Kanzler (Supplementary Note A).



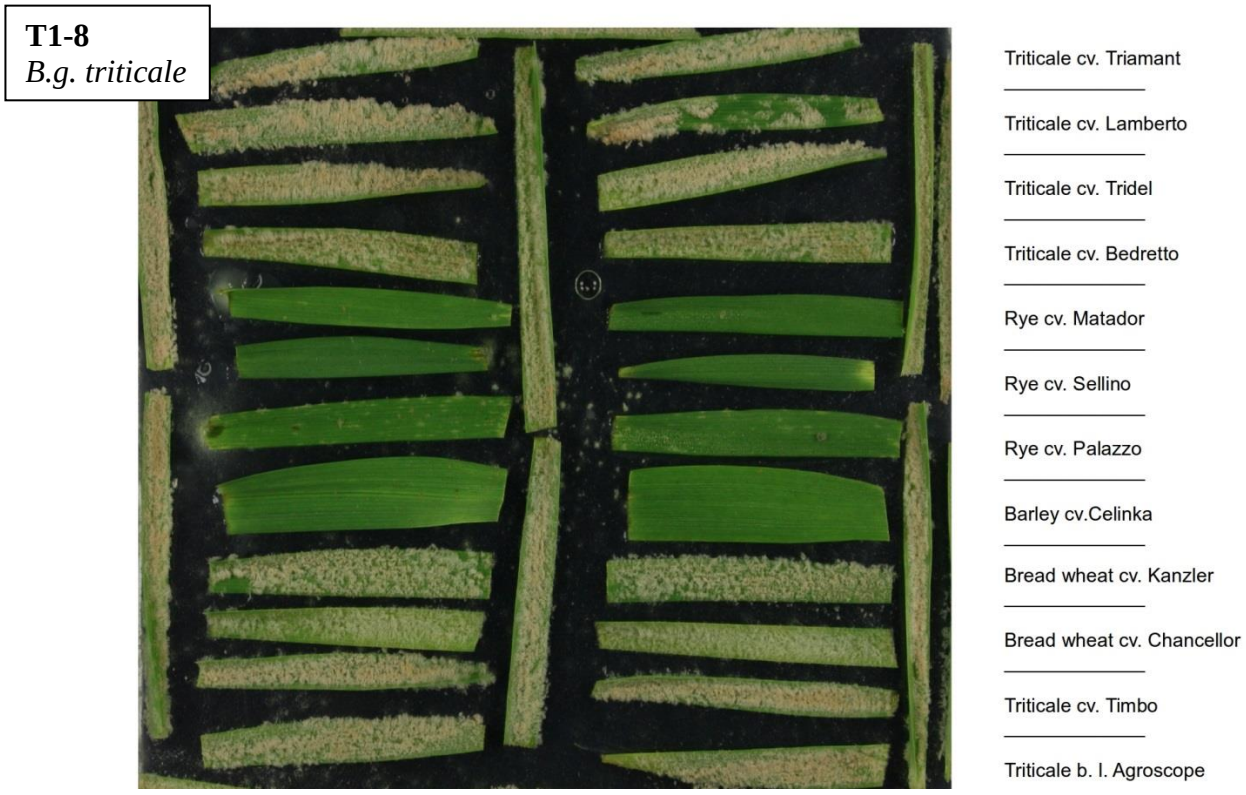
Supplementary Fig. 2 (caption after Supplementary Fig. 23)

Supplementary Fig. 3 (caption after Supplementary Fig. 23)

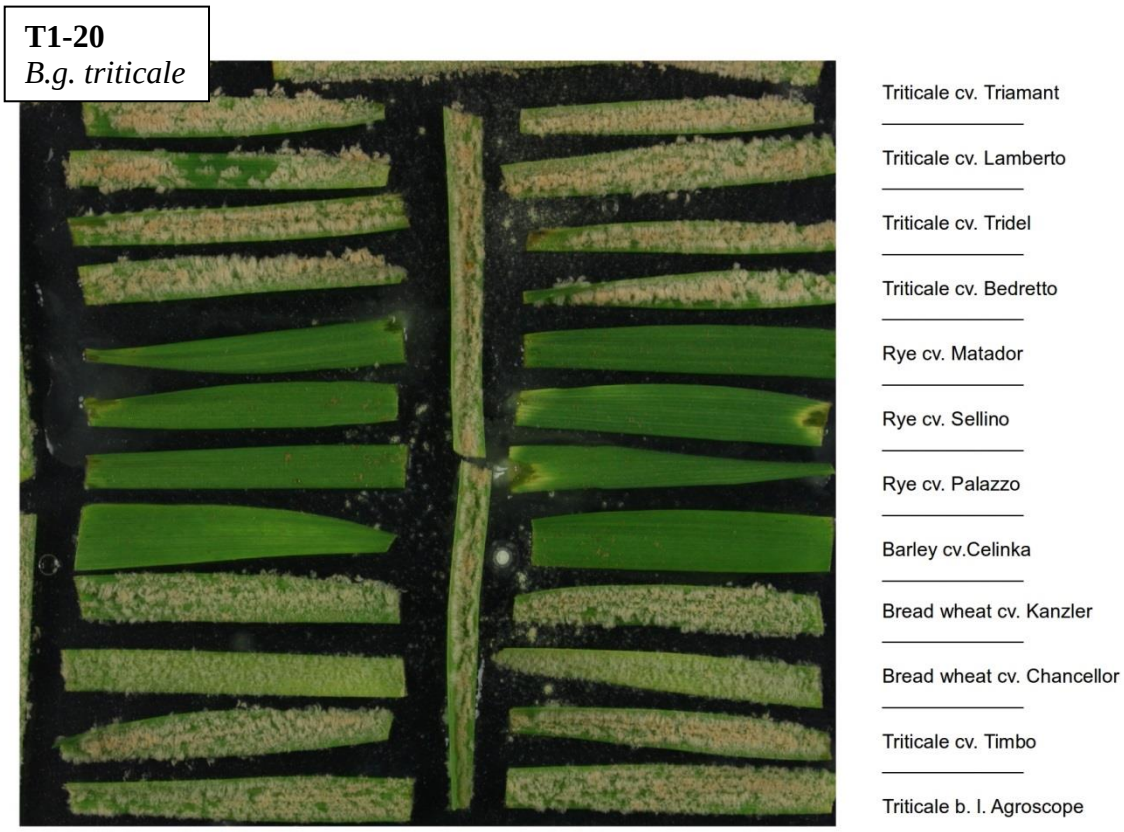
Supplementary Fig. 4 (caption after Supplementary Fig. 23)

Supplementary Fig. 5 (caption after Supplementary Fig. 23)

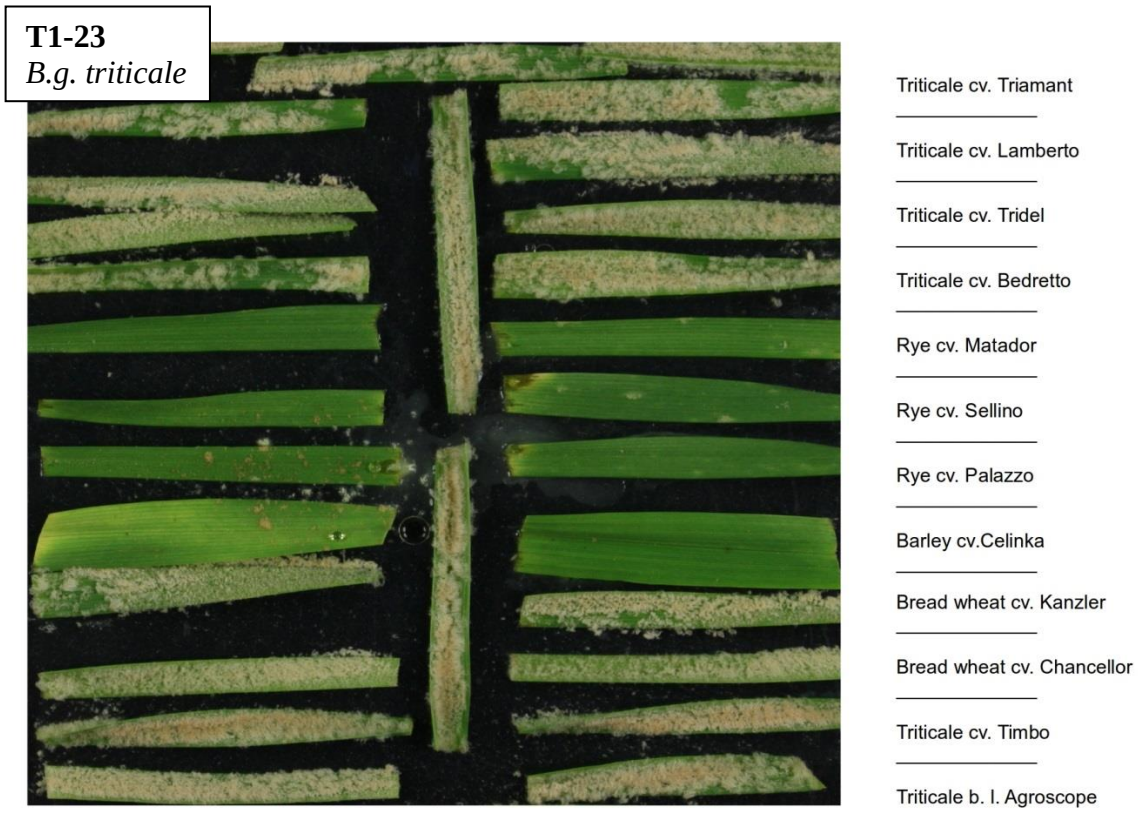
Supplementary Fig. 6 (caption after Supplementary Fig. 23)



Supplementary Fig. 7 (caption after Supplementary Fig. 23)

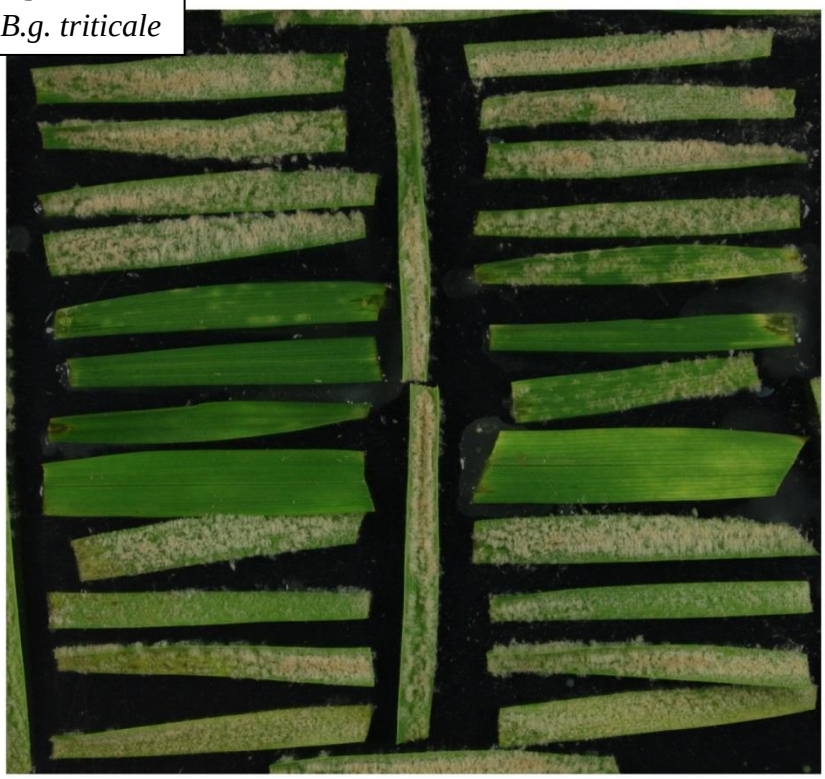


Supplementary Fig. 8 (caption after Supplementary Fig. 23)



Supplementary Fig. 9 (caption after Supplementary Fig. 23)

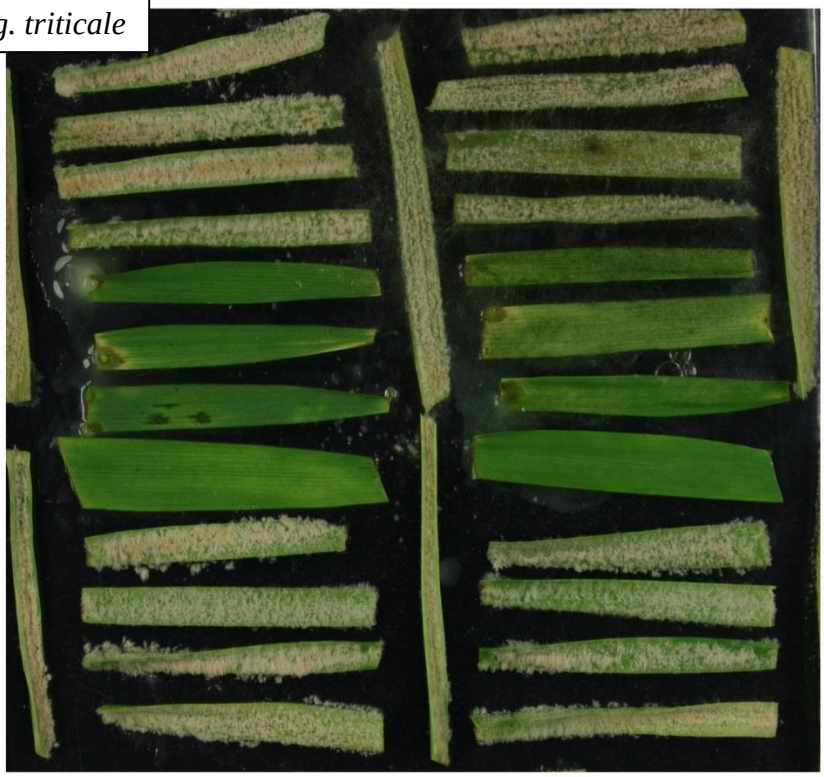
T3-4
B.g. triticale



Triticale cv. Triamant
Triticale cv. Lamberto
Triticale cv. Tridel
Triticale cv. Bedretto
Rye cv. Matador
Rye cv. Sellino
Rye cv. Palazzo
Barley cv.Celinka
Bread wheat cv. Kanzler
Bread wheat cv. Chancellor
Triticale cv. Timbo
Triticale b. l. Agroscope

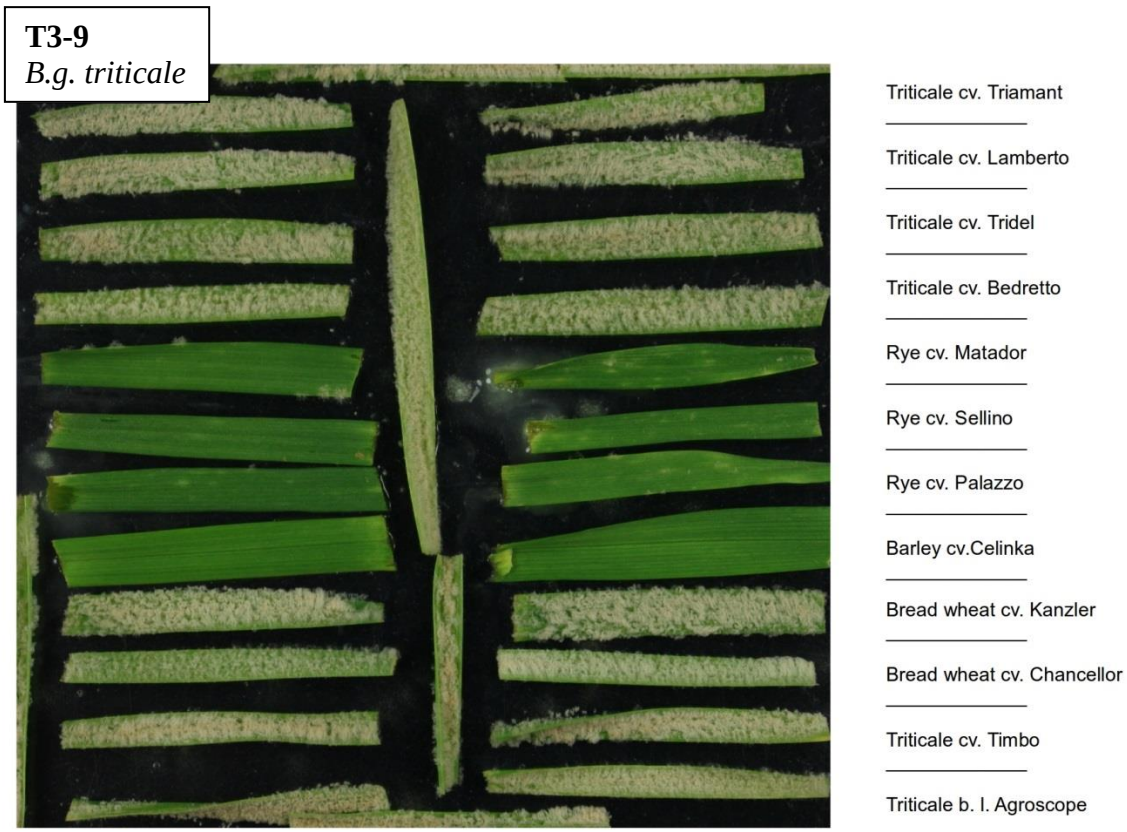
Supplementary Fig. 10 (caption after Supplementary Fig. 23)

T3-8
B.g. triticale

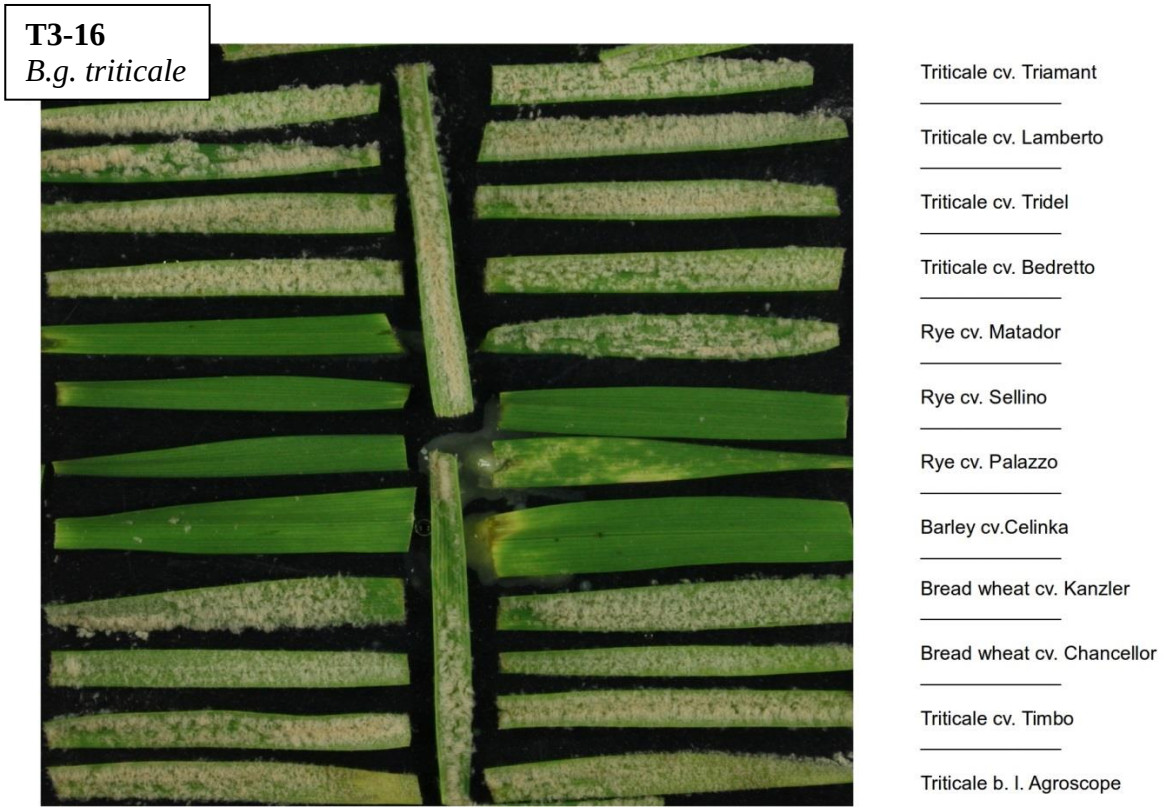


Triticale cv. Triamant
Triticale cv. Lamberto
Triticale cv. Tridel
Triticale cv. Bedretto
Rye cv. Matador
Rye cv. Sellino
Rye cv. Palazzo
Barley cv.Celinka
Bread wheat cv. Kanzler
Bread wheat cv. Chancellor
Triticale cv. Timbo
Triticale b. l. Agroscope

Supplementary Fig. 11 (caption after Supplementary Fig. 23)

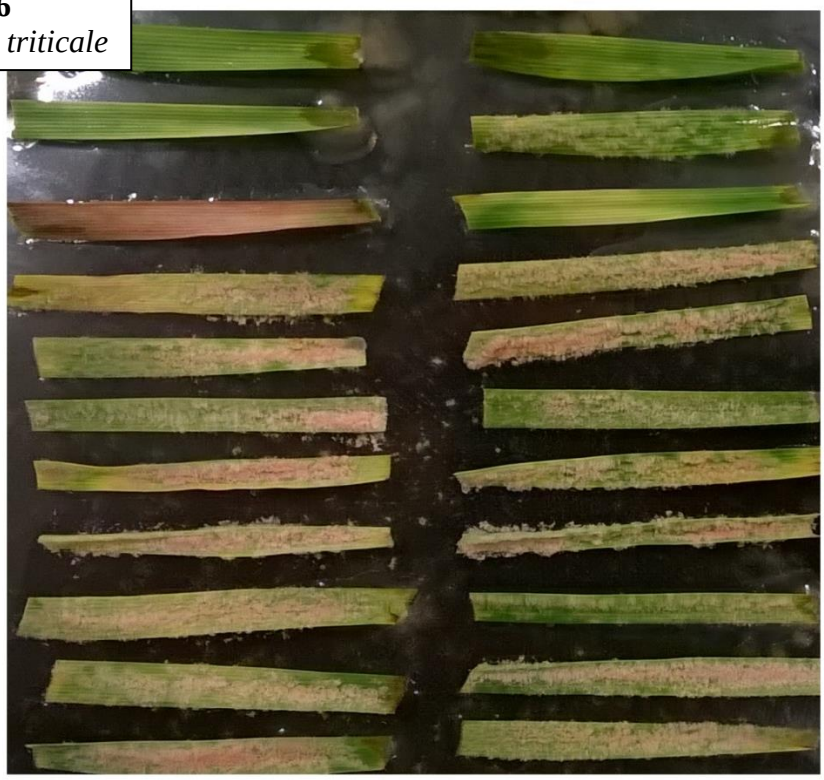


Supplementary Fig. 12 (caption after Supplementary Fig. 23)



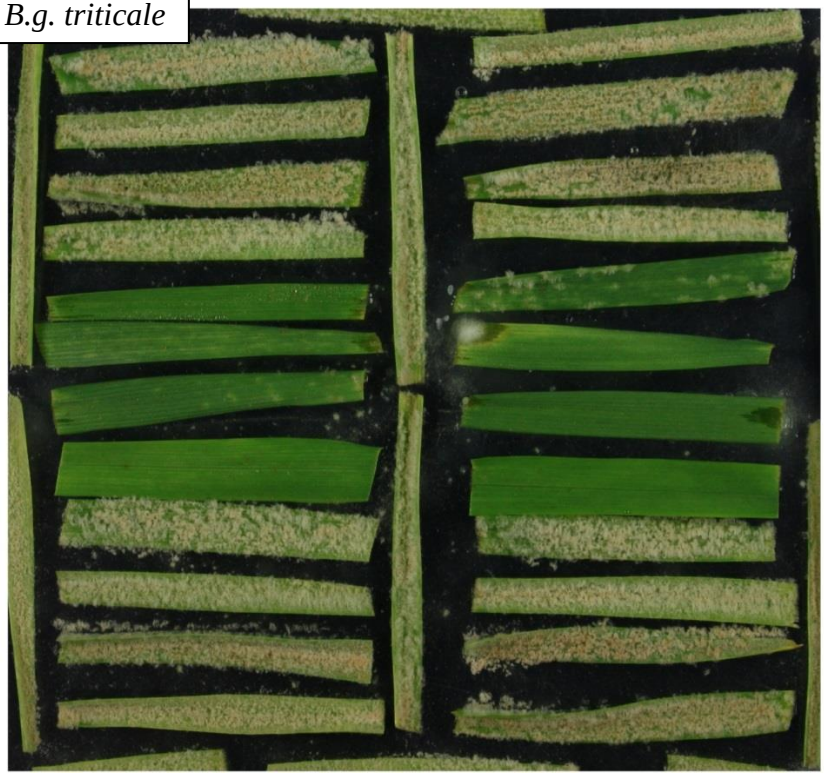
Supplementary Fig. 13 (caption after Supplementary Fig. 23)

T4-6
B.g. triticale



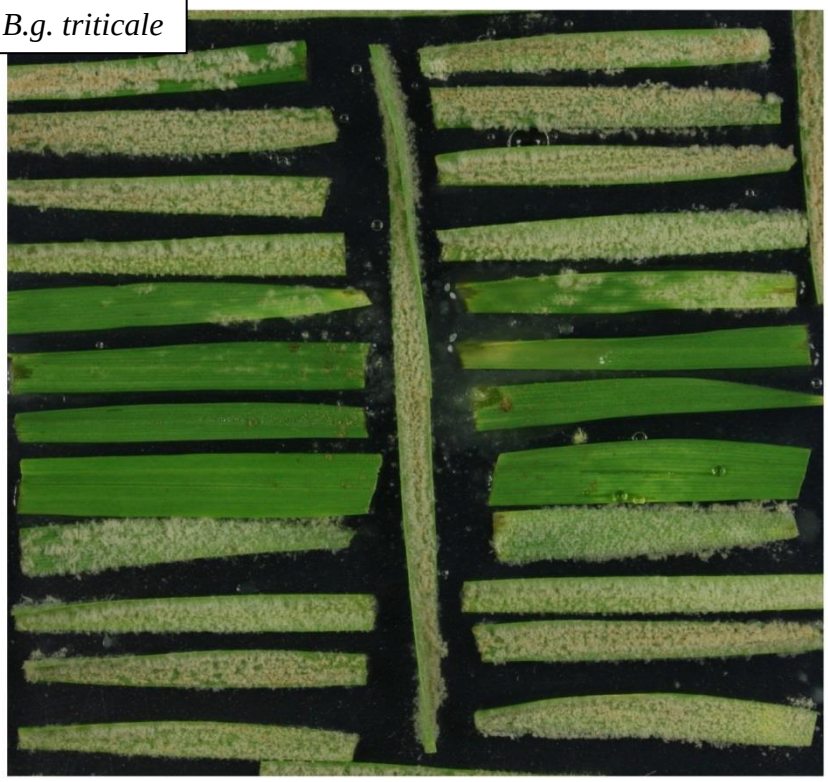
Supplementary Fig. 14 (caption after Supplementary Fig. 23)

T4-7
B.g. triticale



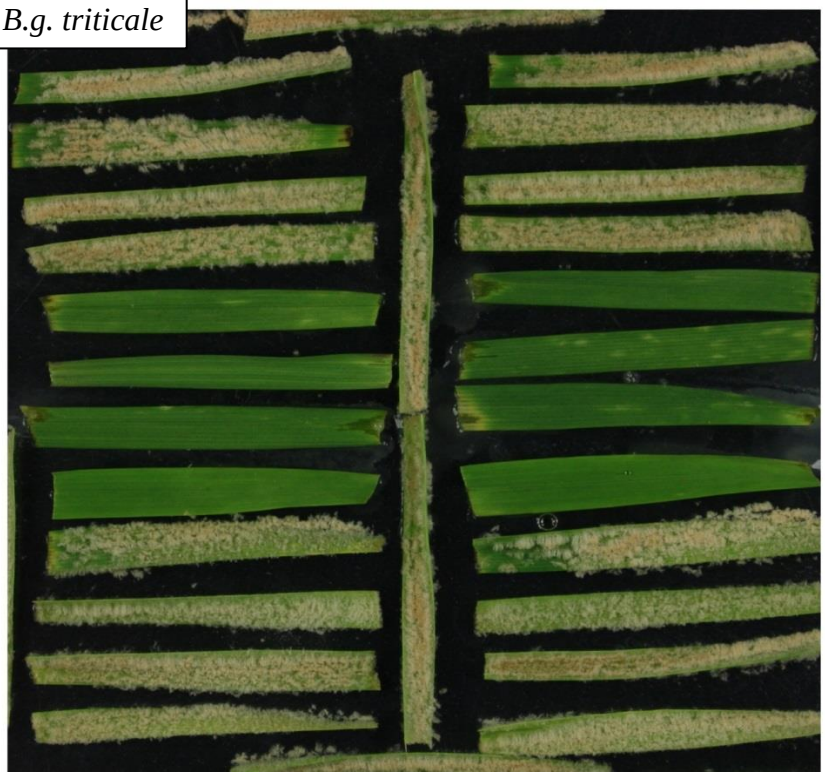
Supplementary Fig. 15 (caption after Supplementary Fig. 23)

T4-19
B.g. triticale



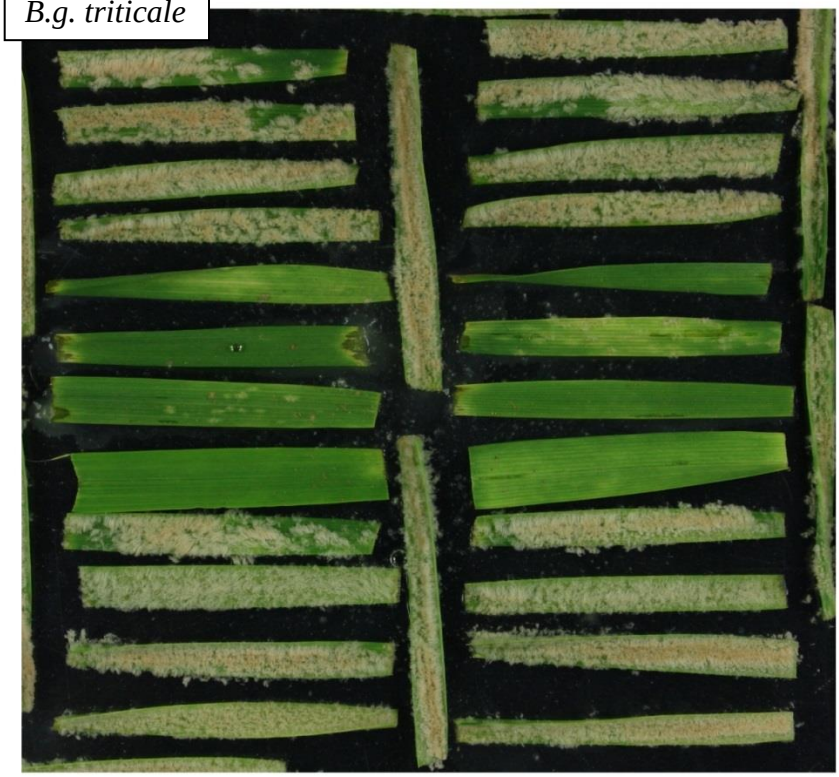
Supplementary Fig. 16 (caption after Supplementary Fig. 23)

T4-20
B.g. triticale



Supplementary Fig. 17 (caption after Supplementary Fig. 23)

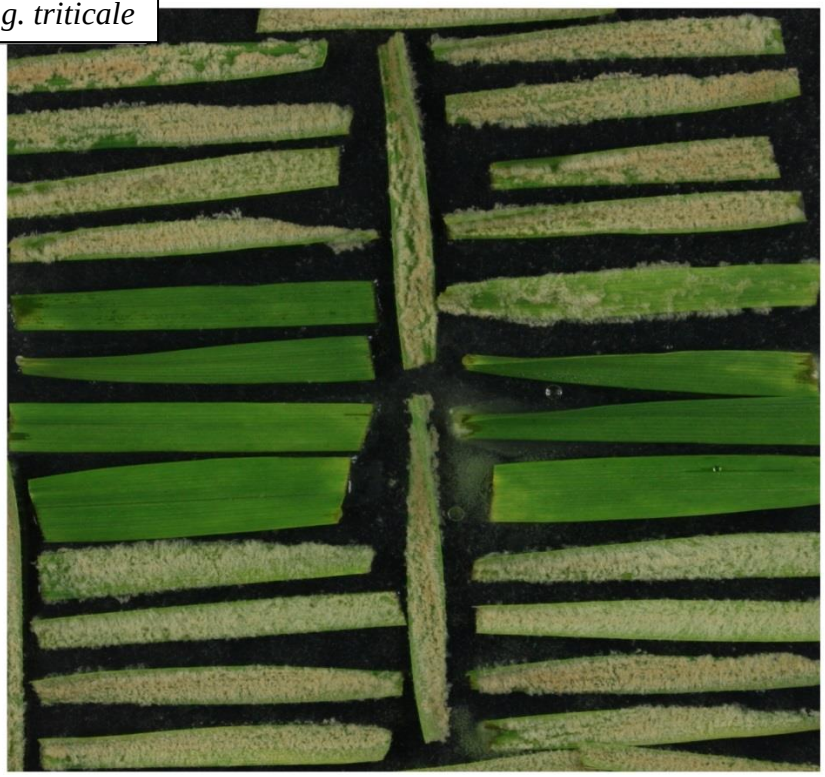
T5-9
B.g. triticale



Triticale cv. Triamant
Triticale cv. Lamberto
Triticale cv. Tridel
Triticale cv. Bedretto
Rye cv. Matador
Rye cv. Sellino
Rye cv. Palazzo
Barley cv.Celinka
Bread wheat cv. Kanzler
Bread wheat cv. Chancellor
Triticale cv. Timbo
Triticale b. l. Agroscope

Supplementary Fig. 18 (caption after Supplementary Fig. 23)

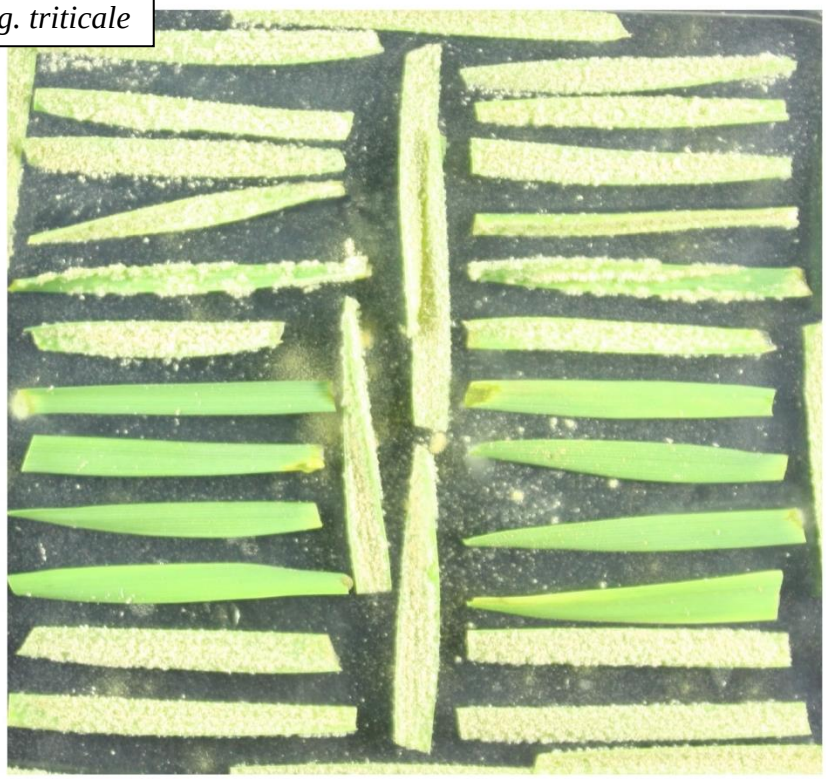
T5-12
B.g. triticale



Triticale cv. Triamant
Triticale cv. Lamberto
Triticale cv. Tridel
Triticale cv. Bedretto
Rye cv. Matador
Rye cv. Sellino
Rye cv. Palazzo
Barley cv.Celinka
Bread wheat cv. Kanzler
Bread wheat cv. Chancellor
Triticale cv. Timbo
Triticale b. l. Agroscope

Supplementary Fig. 19 (caption after Supplementary Fig. 23)

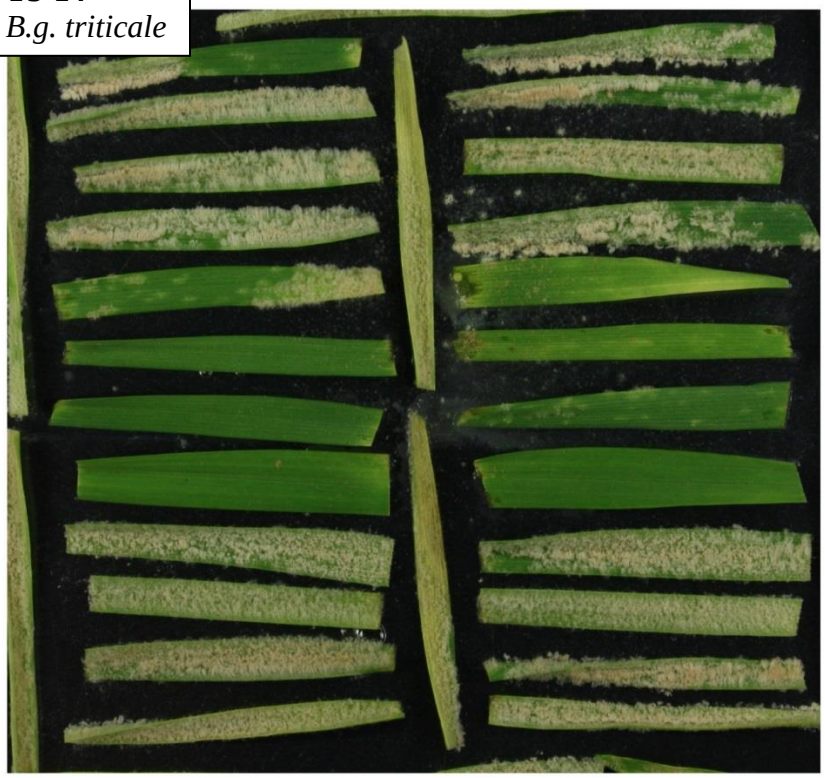
T5-13
B.g. triticale



Triticale cv. Triamant
Triticale cv. Tridel
Triticale cv. Lamberto
Triticale b. l. Agroscope
Triticale cv. Timbo
Triticale cv. Bedretto
Rye cv. Sellino
Rye cv. Palazzo
Rye cv. Matador
Barley cv.Celinka
Bread wheat cv. Chancellor
Bread wheat cv. Kanzler

Supplementary Fig. 20 (caption after Supplementary Fig. 23)

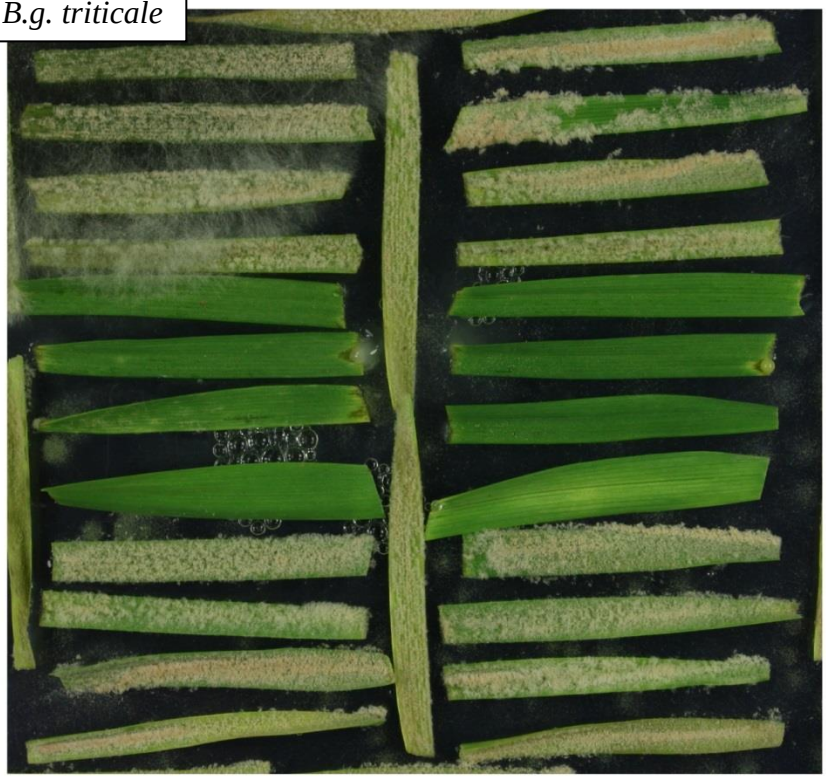
T5-14
B.g. triticale



Triticale cv. Triamant
Triticale cv. Lamberto
Triticale cv. Tridel
Triticale cv. Bedretto
Rye cv. Matador
Rye cv. Sellino
Rye cv. Palazzo
Barley cv.Celinka
Bread wheat cv. Kanzler
Bread wheat cv. Chancellor
Triticale cv. Timbo
Triticale b. l. Agroscope

Supplementary Fig. 21 (caption after Supplementary Fig. 23)

T6-6
B.g. triticale



Triticale cv. Triamant
Triticale cv. Lamberto
Triticale cv. Tridel
Triticale cv. Bedretto
Rye cv. Matador
Rye cv. Sellino
Rye cv. Palazzo
Barley cv.Celinka
Bread wheat cv. Kanzler
Bread wheat cv. Chancellor
Triticale cv. Timbo
Triticale b. l. Agroscope

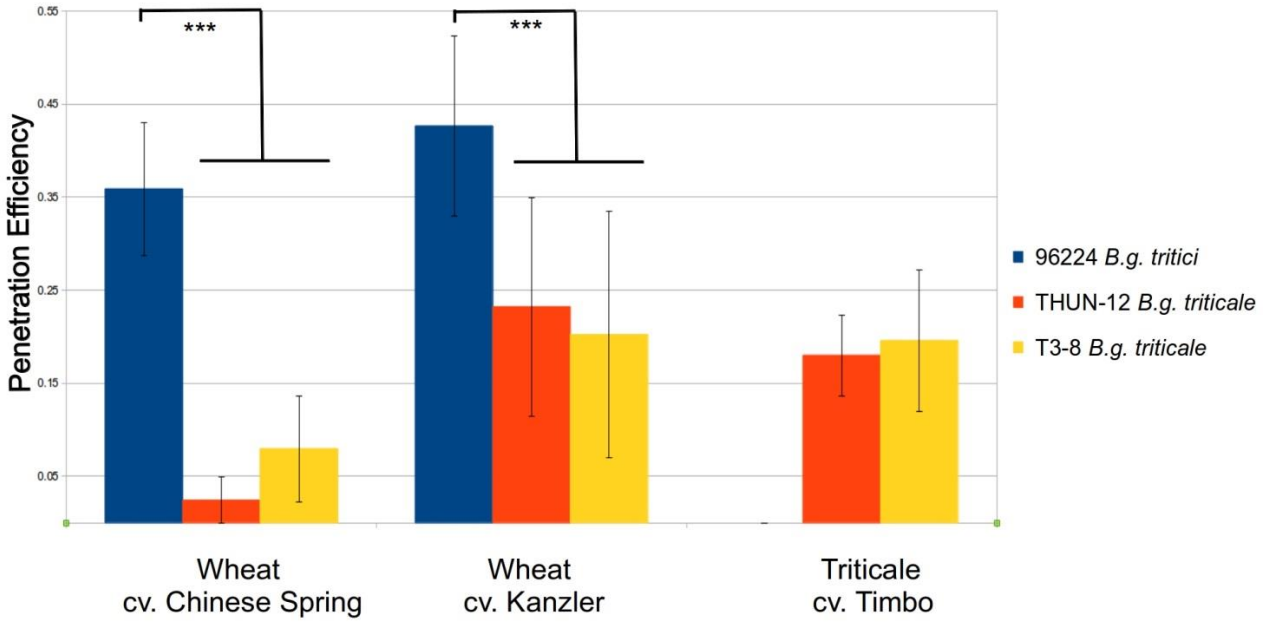
Supplementary Fig. 22 (caption after Supplementary Fig. 23)

THUN-12
B.g. triticale

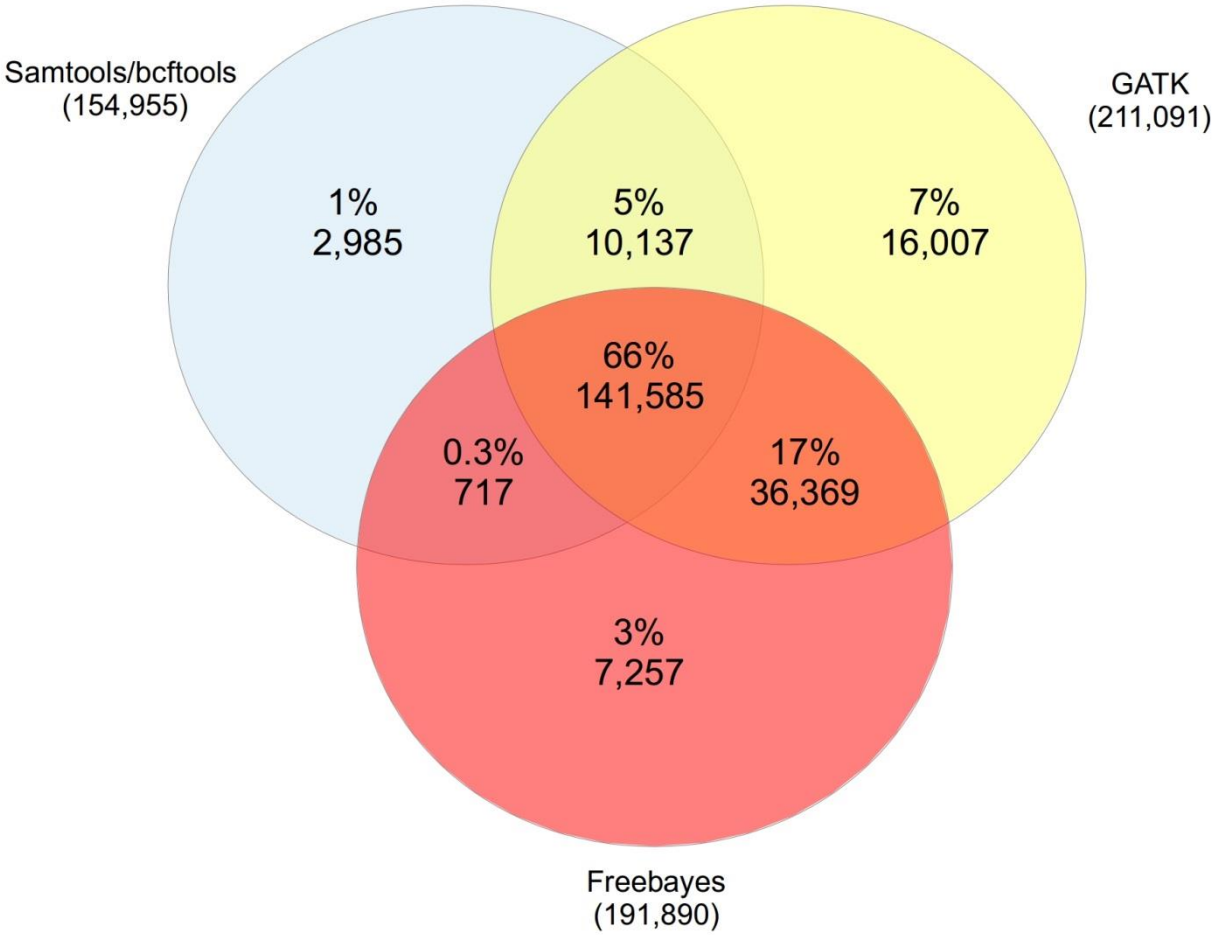
Rye cv. Matador
Rye cv. Palazzo
Rye cv. Sellino
Bread wheat cv. Kanzler
Bread wheat cv. Ch. spring
Bread wheat cv. Chancellor
Triticale b. l. Agroscope
Triticale cv. Timbo
Triticale cv. Bedretto
Pasta wheat cv. Inbar
Triticale cv. Tridel

Supplementary Fig. 23 (caption in the following page)

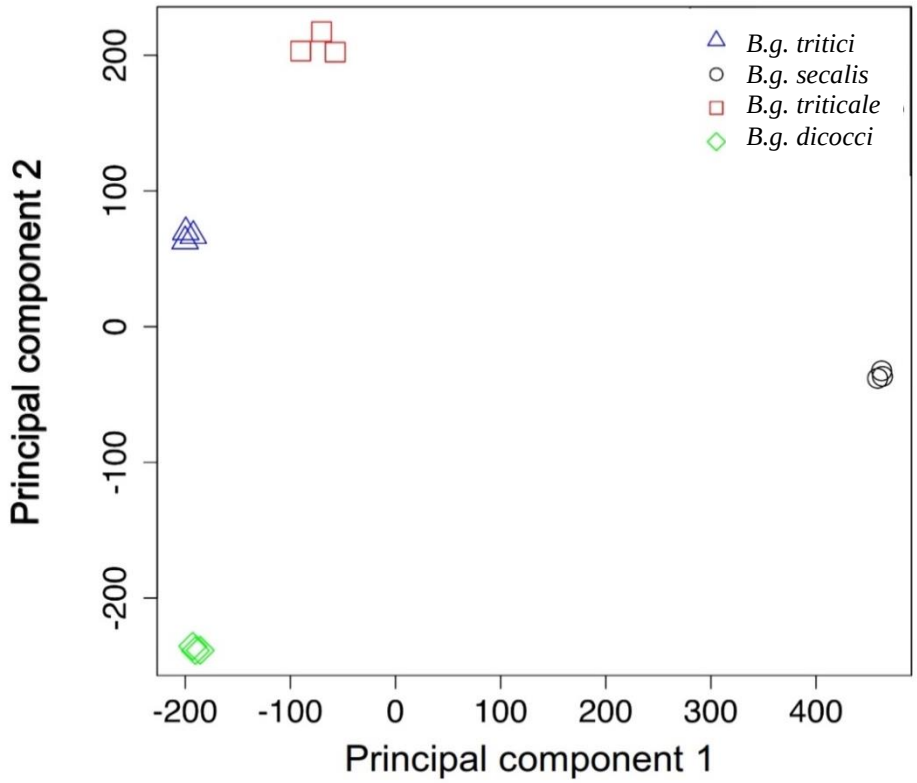
Supplementary Figs. 2-23. Host specificity tests of *B.g. triticale* isolates on rye, tetraploid wheat (durum wheat), hexaploid wheat (bread wheat), triticale and barley (Supplementary Note A). The two columns are two biological replicates of the same set of cultivars (cv.), the name of the plant cultivar is reported on the right, Agroscope is a breeding line (b. l.) obtained from Agroscope Nyon, Switzerland. The names of the isolates are reported in the box at the top of the figure. The results of the host specificity tests are summarized in Supplementary Table 2.



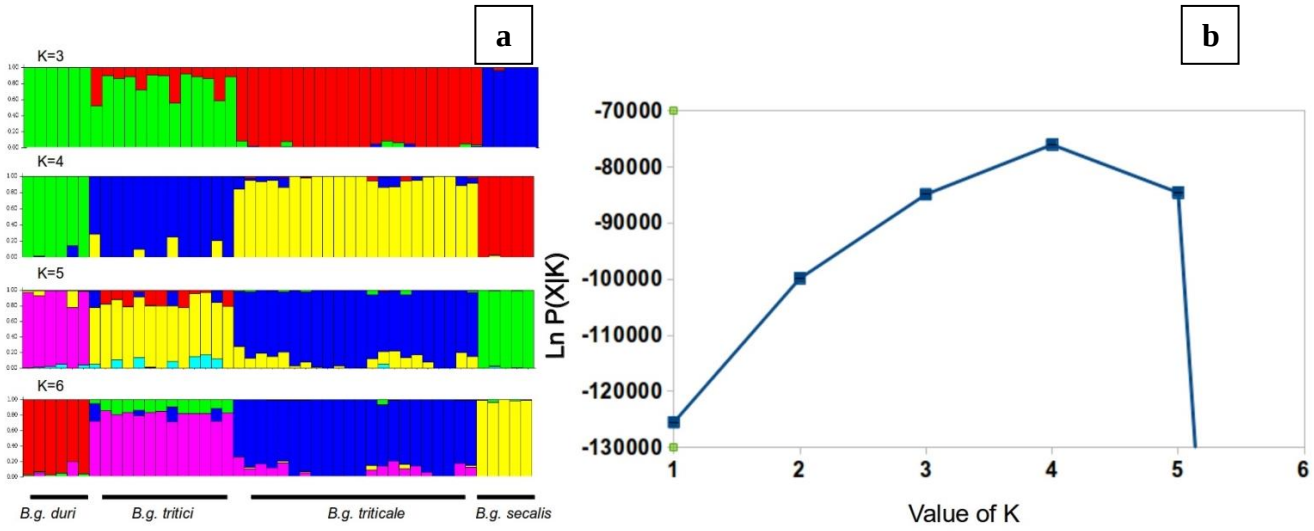
Supplementary Fig. 24. Penetration efficiency of three isolates (96224 in blue *B.g. tritici*, THUN-12 in red and T3-8 in yellow *B.g. triticale*) on two cultivars of bread wheat (Chinese Spring and Kanzler) and one variety of triticale (Timbo) at two days post infection (error bars represent standard deviation between 4 biological replicates). The penetration efficiency is defined as the proportion of successful infection attempts. It is calculated by dividing the number of interactions that results in haustorial formation by the total number of direct interactions (Supplementary Note B). For each plant/pathogen combination we counted at least 50 direct interactions on 4 different leaves (total > 200). Overall *B.g. triticale* isolates have lower penetration efficiency than the *B.g. tritici* isolate for both cultivars of wheat (chi-square p-value < 0.001, all replicates pooled together). Compared to *B.g. tritici* on wheat, the two *B.g. triticale* isolates have a lower penetration efficiency on triticale. Interestingly all isolates have a higher penetration efficiency on the wheat cultivar Kanzler than on Chinese Spring, indicating a higher basal defense level of the latter. It is important to note that these observed differences in penetration efficiency (the first step of infection) do not transmit to the overall macroscopic infection success: after ten days, all the leaf segments of the interactions shown in the figure were fully covered by mildew (with the exception of the *B.g. tritici* isolate 96224 on triticale).



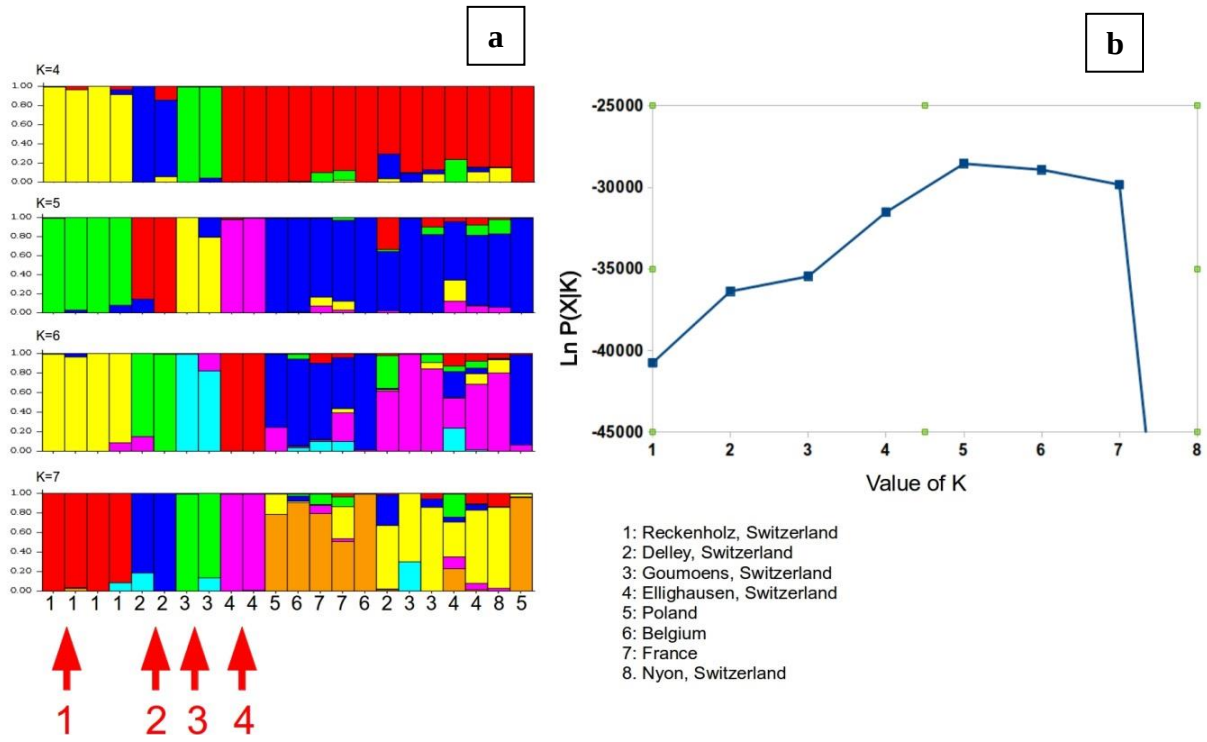
Supplementary Fig. 25. Result of different SNP call pipelines. We tested the quality of our SNP call done with samtools (Li et al. 2009) by comparing it with the results of two alternative approaches (GATK (DePristo et al. 2011), Freebayes (Garrison and Marth 2012)) on 12 isolates (Supplementary Note D). Samtools variants were filtered by coverage (> 20 X), mapping quality (> 20) and proportion of reads carrying the alternative call (> 0.95). GATK and Freebayes low-confidence variants were filtered out by quality (>20) and coverage (> 20X). The percentages represent proportions of the total number of SNPs called in the 12 isolates by each of the methods. In the overlapping area of the circles, the percentage of SNPs called by two or three methods is indicated. The numbers below percentages are absolute number of SNPs (median between the 12 isolates). The samtools SNP call was the most stringent one as only 9% of SNPs called by samtools were not called by any of the other two pipelines.



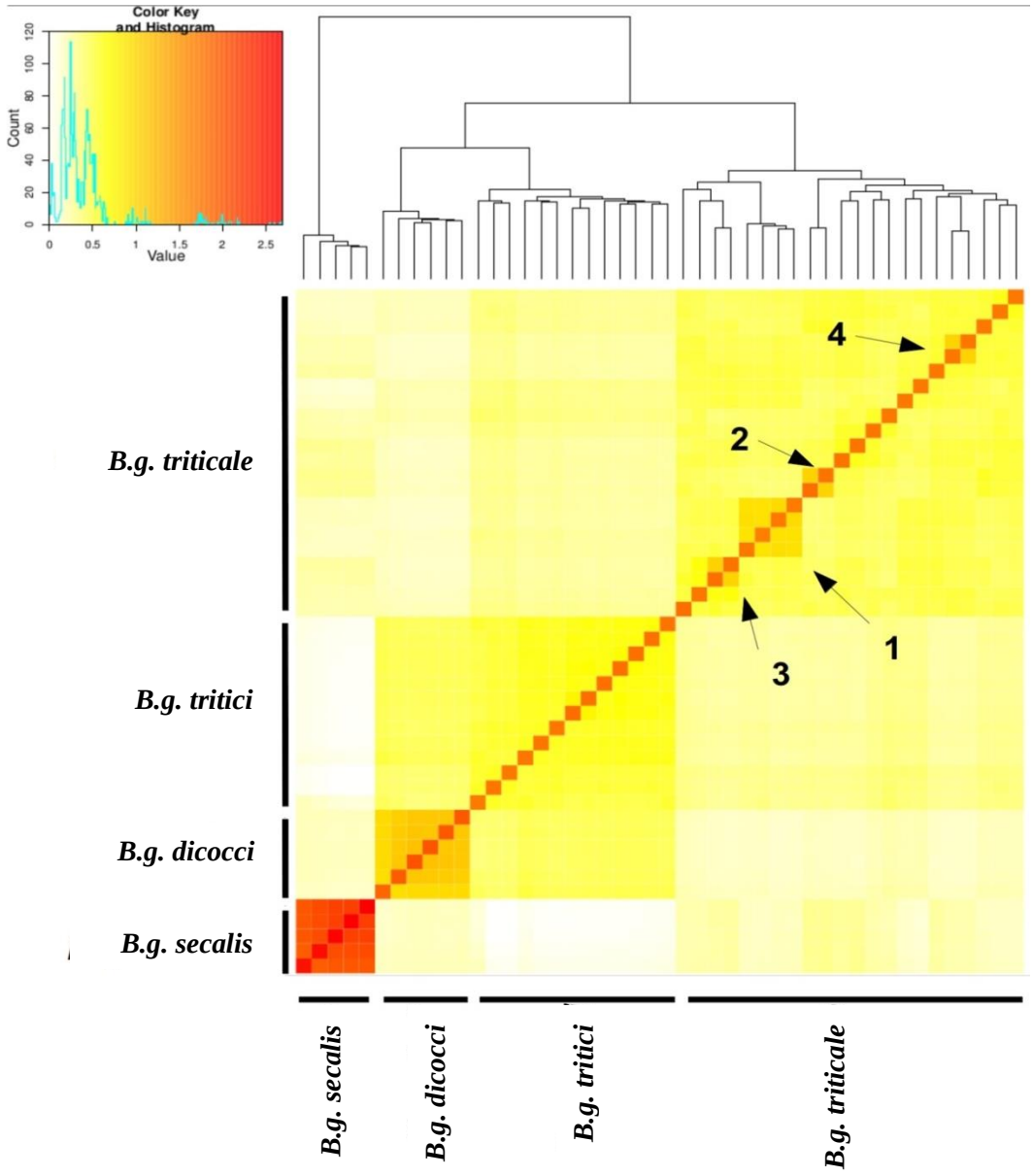
Supplementary Fig. 26. Principal component analysis **(**PCA) of SNPs called with the GATK (DePristo et al. 2011) pipeline. The PCA was performed with the R package adegenet (Jombart and Ahmed 2011). A total of 12 isolates are included (isolates 217, 97 and103 for *B.g. tritici*, S-1201, S-1459, S-1391 for *B.g secalis*, 209, 220 and 58 for *B.g. dicocci*, THUN-12, T4-19 and BAH-2 for *B.g. triticale*). The four *ff. spp.* cluster together as in Fig. 2. This analysis confirmed that a different SNP call pipeline produces very similar results (Supplementary Note D).



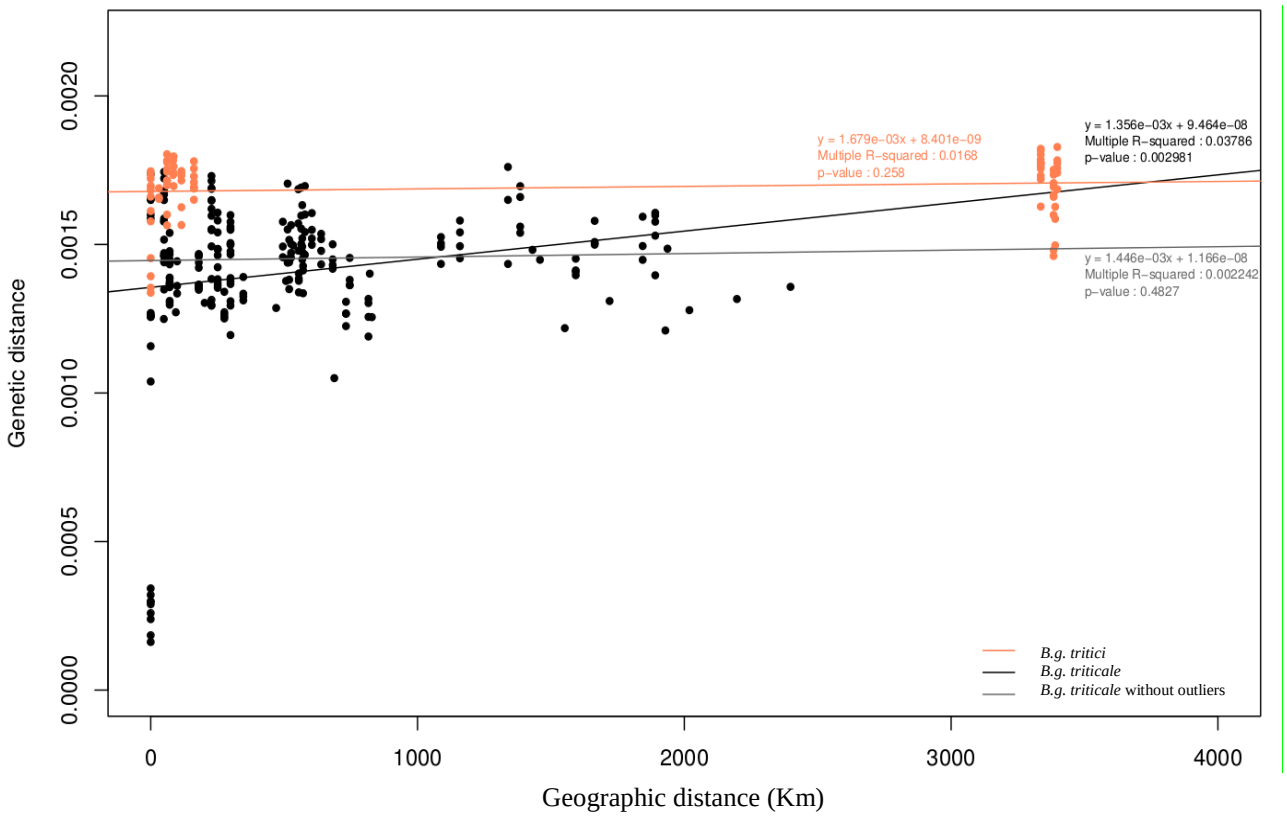
Supplementary Fig. 27. STRUCTURE analysis on all isolates. **a)** Barplots result of STRUCTURE (Pritchard et al. 2000) for different values of K (3 to 6). Each vertical bar represents one isolate, the colors represent the proportion of inferred ancestry from K different ancestral populations. STRUCTURE clearly distinguishes the different *ff. spp*. as different clusters. Some individuals of *B.g. tritici* and *B.g. triticale* are predicted to be admixed. **b)** Plot of the Ln probability of the data for different values of K (1 to 6), the probability is highest for K = 4. The value of the Ln probability of the data for K = 6 is -418,612 and falls out of the graph (Supplementary Note E).



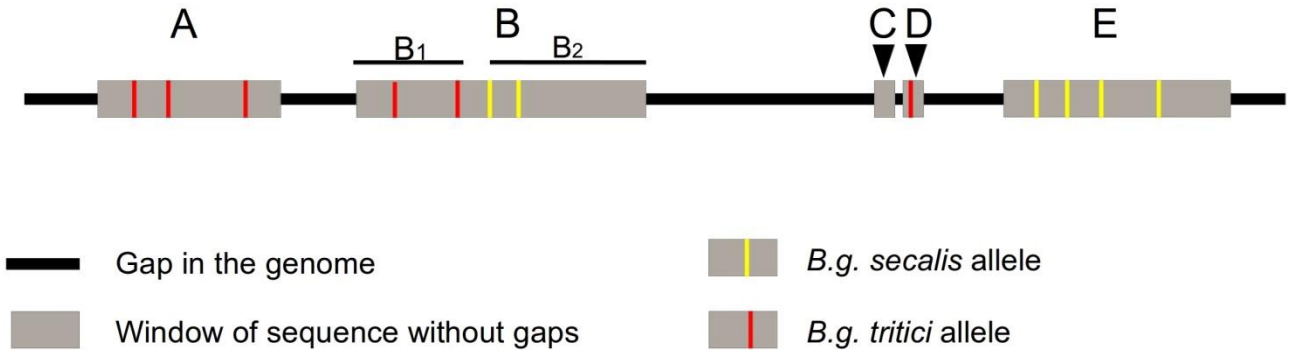
Supplementary Fig. 28. STRUCTURE analysis of *B.g. triticale* (3,861 randomly selected SNPs) For this analysis we ran 500,000 generations plus 100,000 replications as burn-in. **a)** Barplots result of STRUCTURE (Pritchard et al. 2000) for different values of K (4 to 7). Each vertical bar represents one isolate. The colors represent the proportion of inferred ancestry from K different ancestral populations. STRUCTURE recognizes four different groups of isolates (red arrows and numbers): groups 1, 2, 3 and 4 are composed of isolates that have been collected the same day in the same field (in four different Swiss locations: Reckenholz (1), Delley (2)), Goumoens (3) and Hellighausen (4)). These groups are composed of isolates that are related between them (Supplementary Fig. 29). All the other isolates are in one single group with K =5 but divided in two different groups with K > 5, in one subgroup there are all remaining Swiss isolates, in the other isolates from France, Poland and Belgium. Black numbers indicate locality of sampling for each isolate. **b)** Plot of the Ln probability of the data for different values of K (1 to 8), the probability is highest for K = 5. The value of the Ln probability of the data for K = 8 is -74,363 and falls out of the graph (Supplementary Note E).



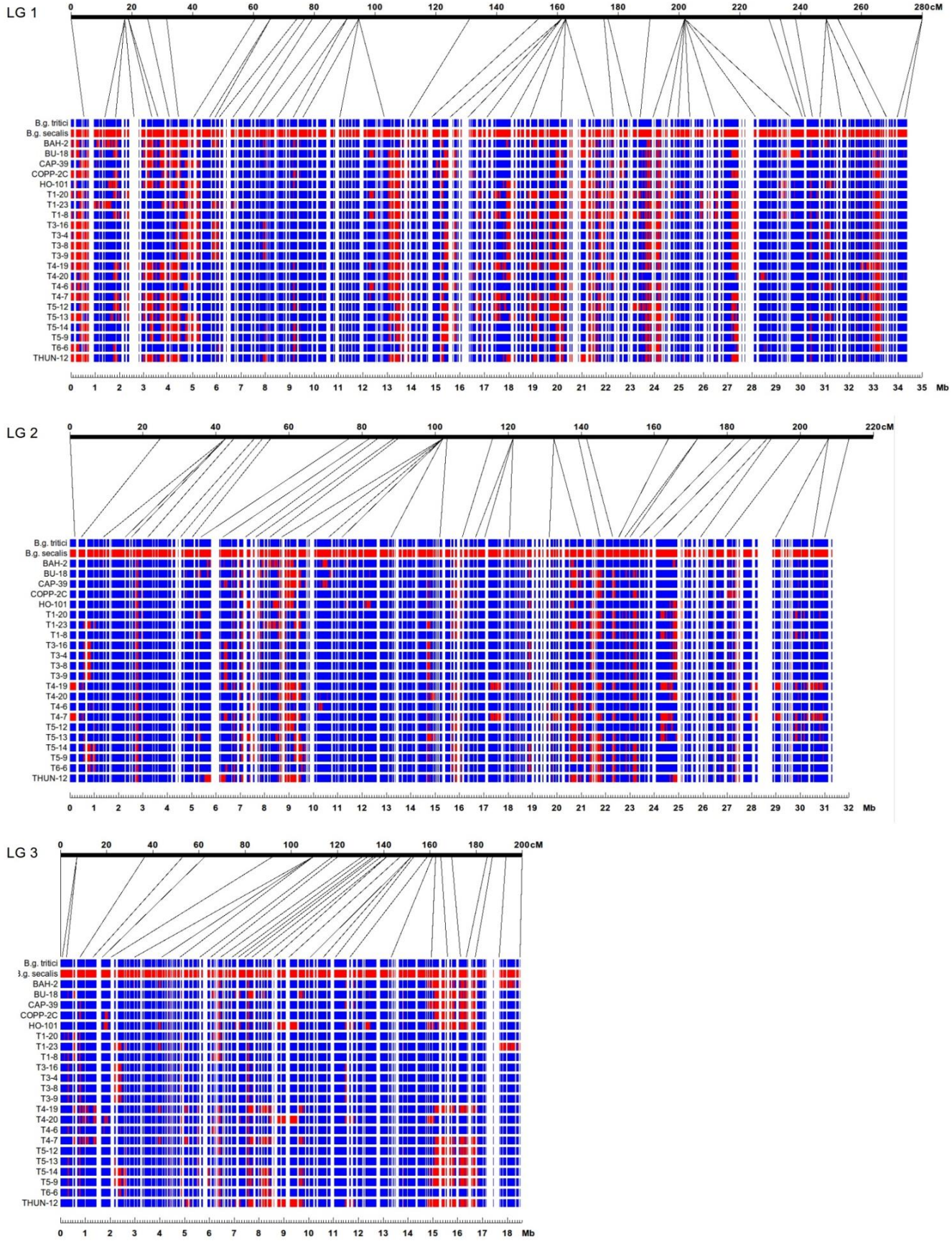
Supplementary Fig. 29. Cluster and Kinship results of GAPIT (Lipka et al. 2012) using all SNPs of the sequenced *B. graminis* isolates present in at least two individuals. The figure shows a heat map of values of the Kinship matrix. Darker colors indicate that two isolates are more likely to be identical by descent (to have the same genotype because they are related). The different *ff. spp.* cluster in 4 different groups. *B.g. secalis* and *dicocci* isolates are more related between them than *B.g. tritici* and *triticale* isolates. Moreover in *B.g. triticale* there are five groups of isolates that have a higher probability of sharing an allele between them by descent than any other pairwise comparison of *B.g triticale* isolates. These groups (1, 2, 3, 4) are also detected by STRUCTURE's analysis. (Supplementary Note E). This shows that the population structure in *B.g. triticale* detected by STRUCTURE is probably due to a sampling bias: in our sample there are isolates that are related between them, in each group (1 - 4) all isolates have been sampled the same day in the same field.



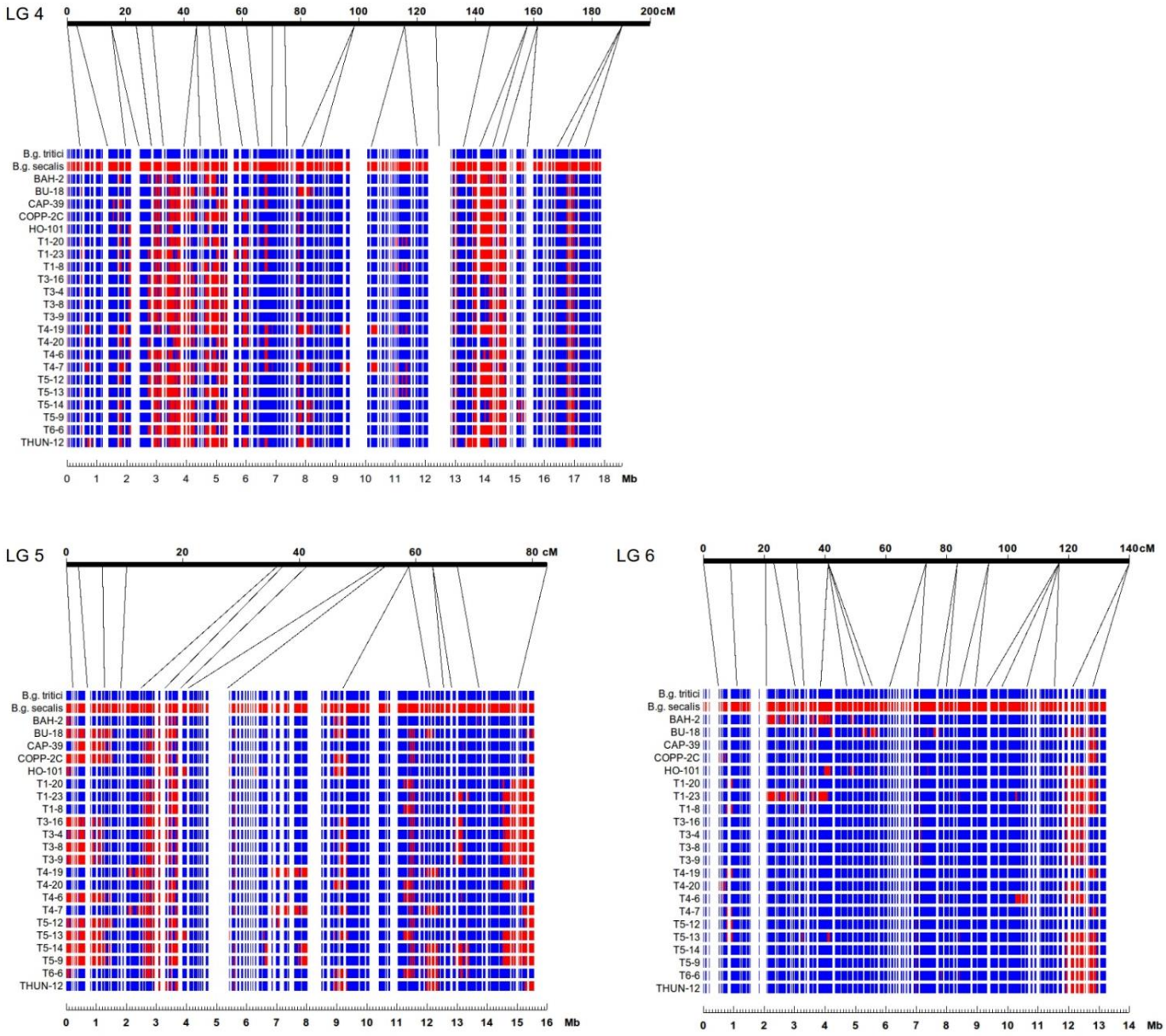
Supplementary Fig. 30. Isolation by distance in *B.g tritici* and *B.g. triticale*. Pairwise genetic distances (SNP per base) and geographic distances (Km) are plotted, (*B.g. tritici* in red, *B.g. triticale* in black). The steepness of the regression lines (coefficient in the equations) indicates the extent of isolation between isolates due to geographic distance. The Multiple R-squared is a measure of the variance fitted by the linear model (0 none, 1 all). The p-value tests the null hypothesis that the coefficient is equal to zero, therefore no effect of geography on genetic distances. *B.g. tritici* does not show a positive correlation between genetic and geographic distance, also *B.g. triticale*, after exclusion of outliers (bottom left in the graph), does not show any isolation by distance (grey line) (Supplementary Note E). These outliers represent all possible pairwise combination of isolates in group 1, 2, 3 and 4 (recognized by STRUCTURE and Kinship matrix) with isolates belonging to the same group.



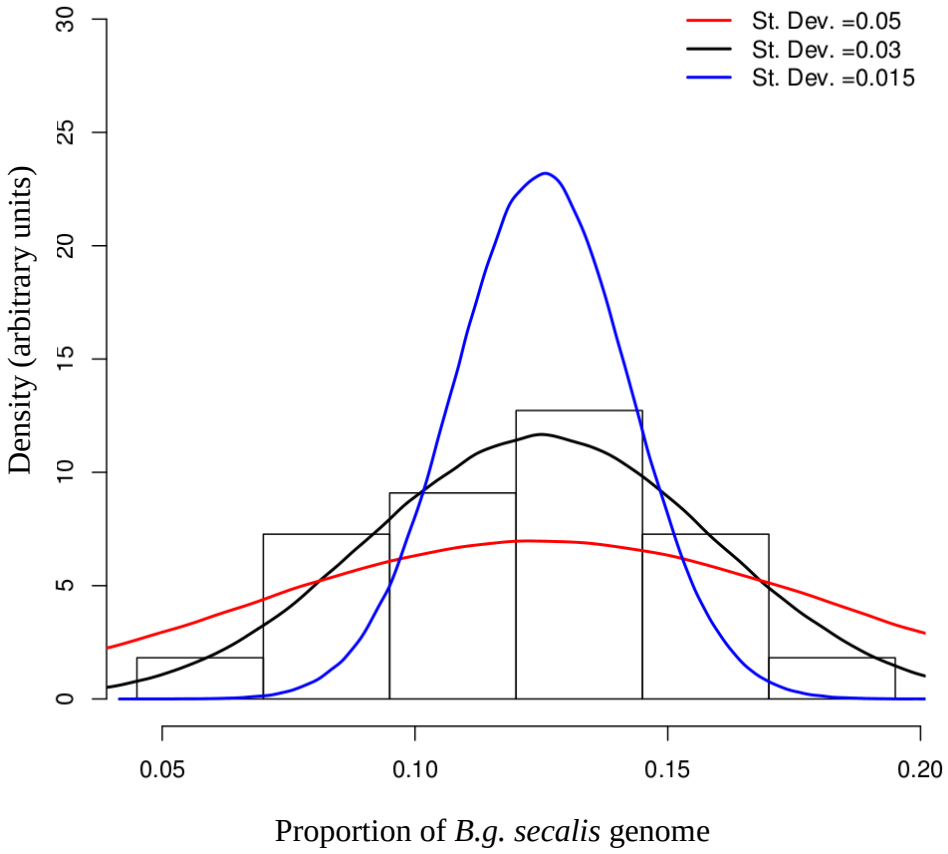
Supplementary Fig. 31. Identification of genomic segments originating from different *formae speciales* in *B.g. triticale*. We divided the *B.g. tritici* reference genome sequence into 22,271 windows that do not contain sequence gaps (average window size: 3,682 bp). Sequence gaps are mainly caused by the very high repeat content of the *B.g. tritici* genome. The characteristics of the different window types and their use in the analyses are described in Supplementary Note G.



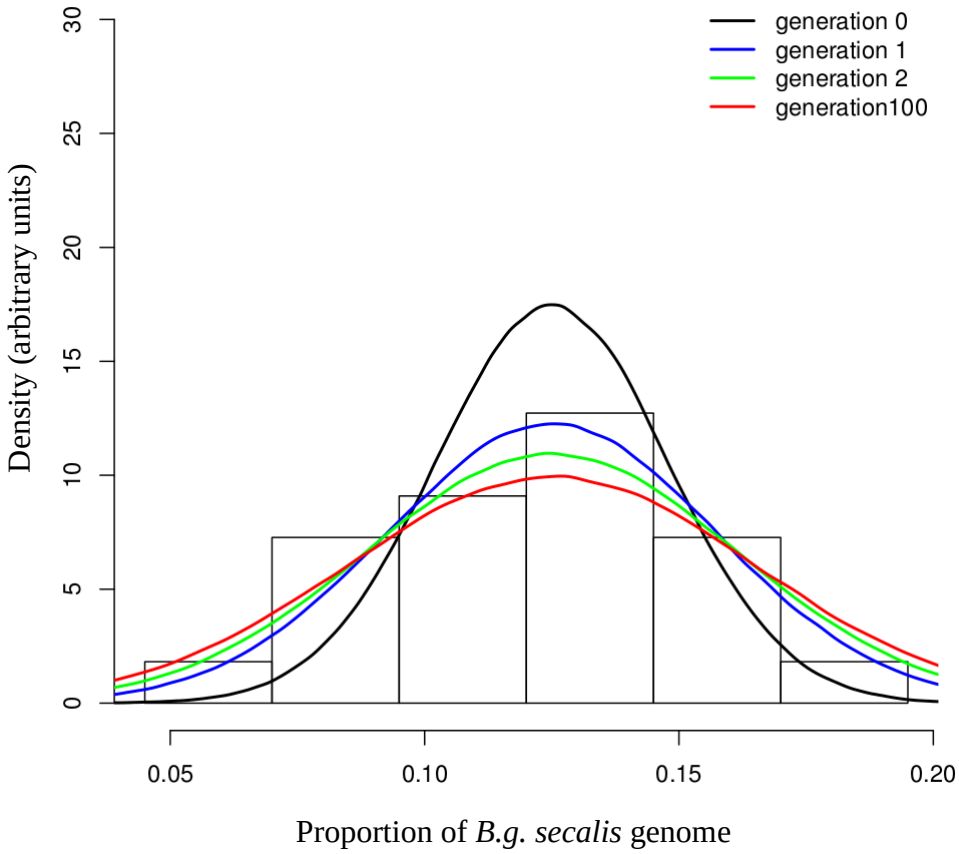
Supplementary Fig. 32 (caption in the following page).



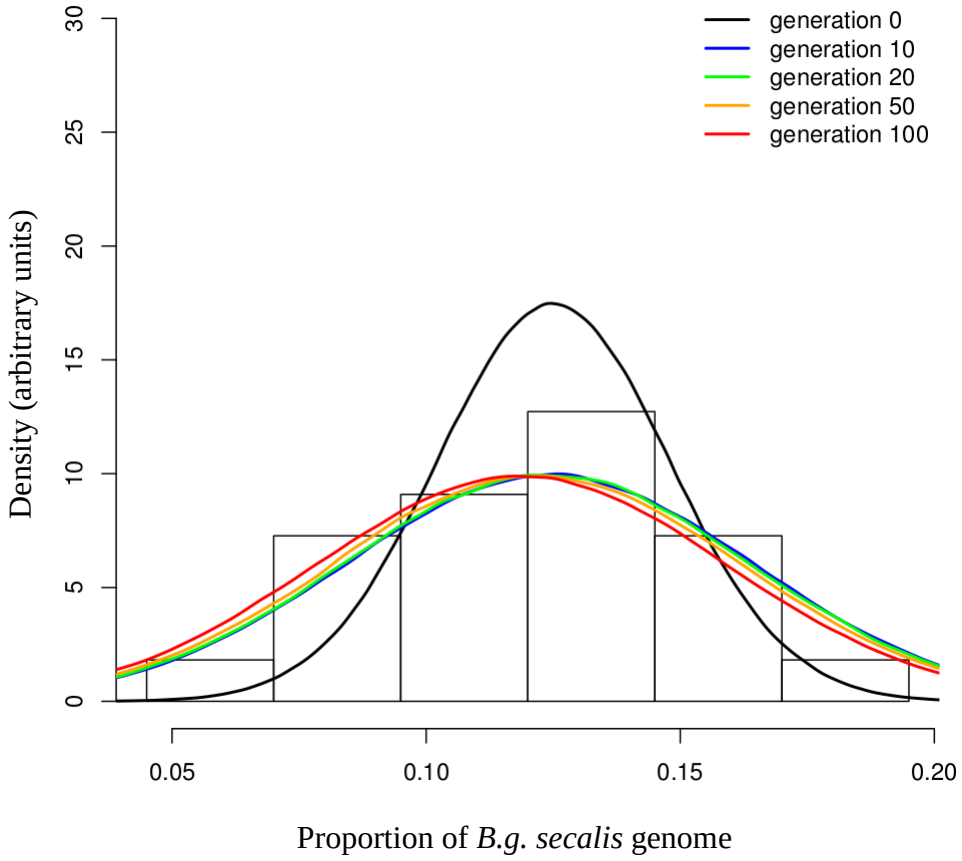
Supplementary Fig. 32 and 33. Substitution pattern on linkage groups of *B.g. triticale*. Nucleotide substitutions in *B.g. triticale* isolates compared to *B.g. tritici* and *B.g. secalis* on the 6 larger linkage groups of *B. graminis*. Fixed polymorphisms specific for *B.g. tritici* isolates are depicted in blue while those specific for *B.g. secalis* isolates are depicted in red. There is a characteristic mosaic pattern that indicates a relatively small number of recombination events in a young hybrid species. The linkage groups are based on the genetic map of *Blumeria graminis* published in Bourras et al. (2015) . For each linkage group we selected one genetic marker and linked the genetic marker to the physical contigs where the markers were derived from (Supplementary Note H).



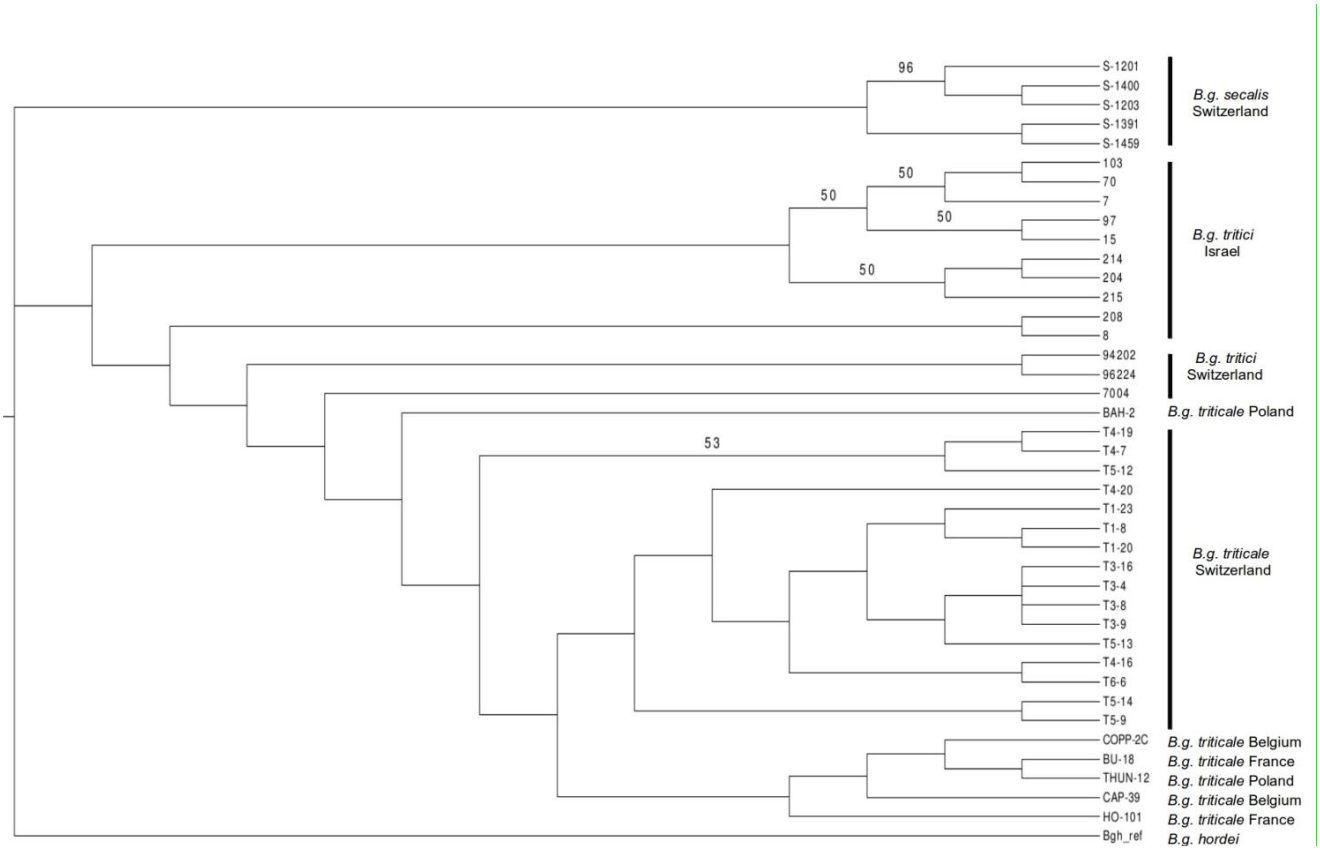
Supplementary Fig. 34. Results of back-crossing simulations: the histogram shows the observed distribution of the contribution of the *B.g. secalis* parent to the genomes of 22 *B.g. triticale* isolates. The three curves represent the result of 10,000 simulations of one hybridization and two back-crosses with different values of standard deviation (Supplementary Note H). Observed and simulated data follow the same distribution.



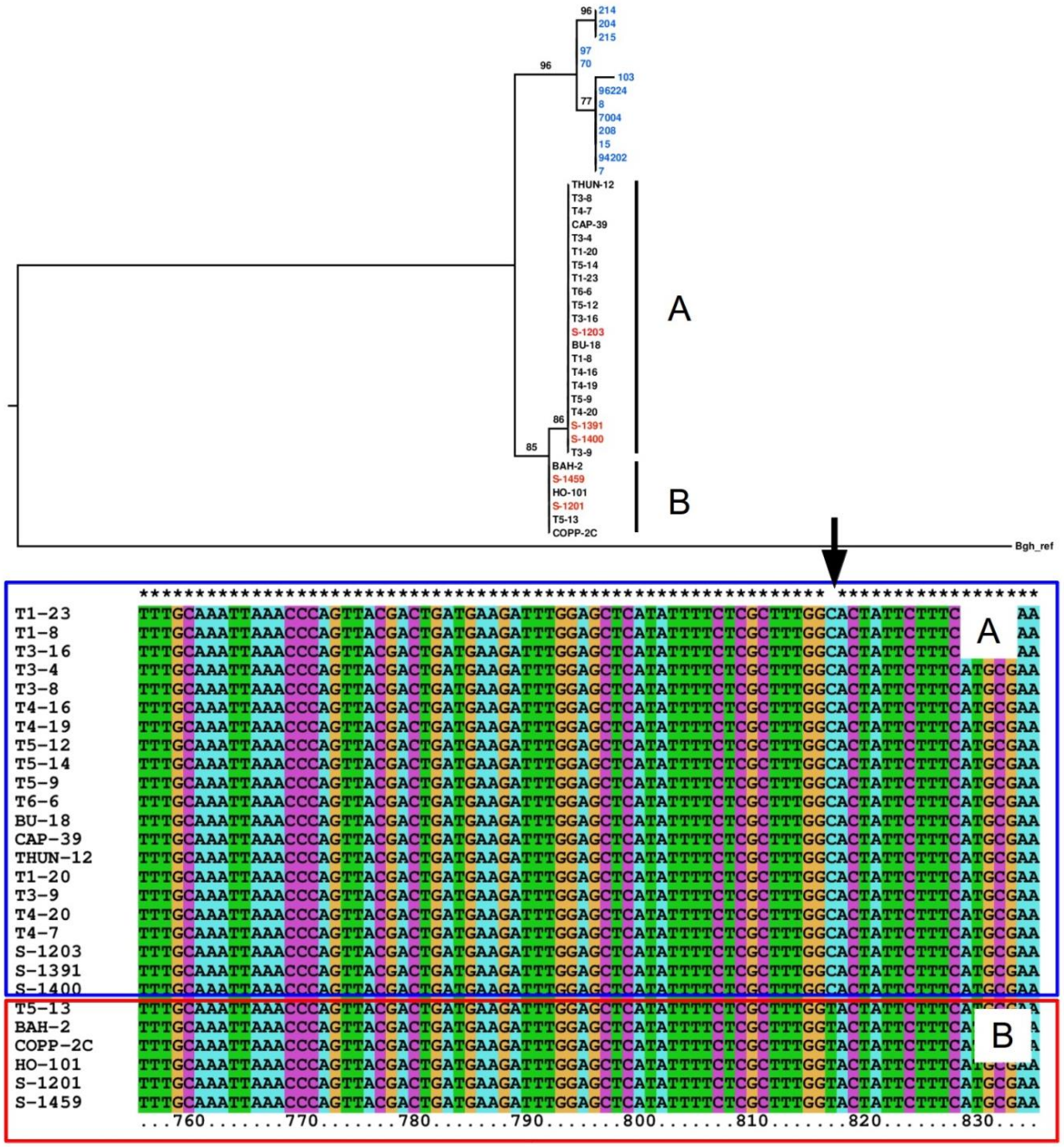
Supplementary Fig. 35. Results of back-crossing simulations: the histogram shows the observed distribution of the contribution of the *B.g. secalis* parent to the genomes of 22 *B.g. triticale* isolates. The four colored lines represent the result of simulations of one hybridization and two back-crosses (standard deviation = 0.002) followed by random mating between *B.g. triticale* isolates (Fisher-Wright population n = 10,000) (Supplementary Note H). With increasing numbers of generations the distributions become flatter. This effect is strong in the first two generations and becomes limited with following generations (e.g. the difference between generation 2 and generation 100 is smaller than between generation 1 and generation 2).



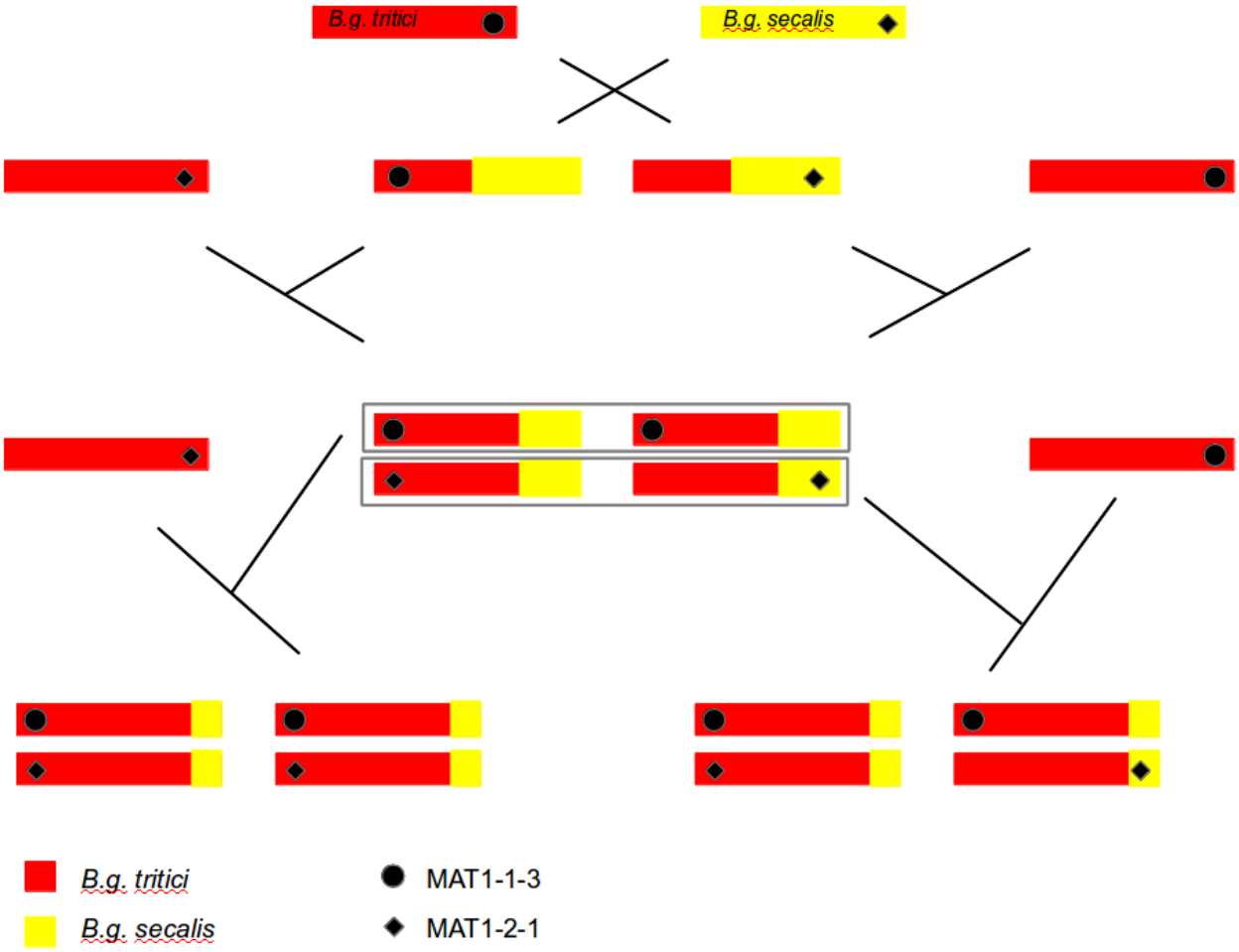
Supplementary Fig. 36. Results of back-crossing simulations: the histogram shows the observed distribution of the contribution of the *B.g. secalis* parent to the genomes of 22 *B.g. triticale* isolates. The colored lines represent the result of simulations of one hybridization and two back-crosses (standard deviation = 0.002) followed by random mating between *B.g. tritical*e isolates (Fisher-Wright population n = 10,000) with limited gene flow (10 isolates have a *B.g. tritici* parent at every generation) (Supplementary Note H). With increasing numbers of generations the distribution shifts to the left. Nevertheless in this simulation after 100 generations observed and simulated data follow the same distribution. Therefore, we cannot exclude the presence of gene flow between *B.g. tritici* and *B.g. triticale* after the origin of the latter.



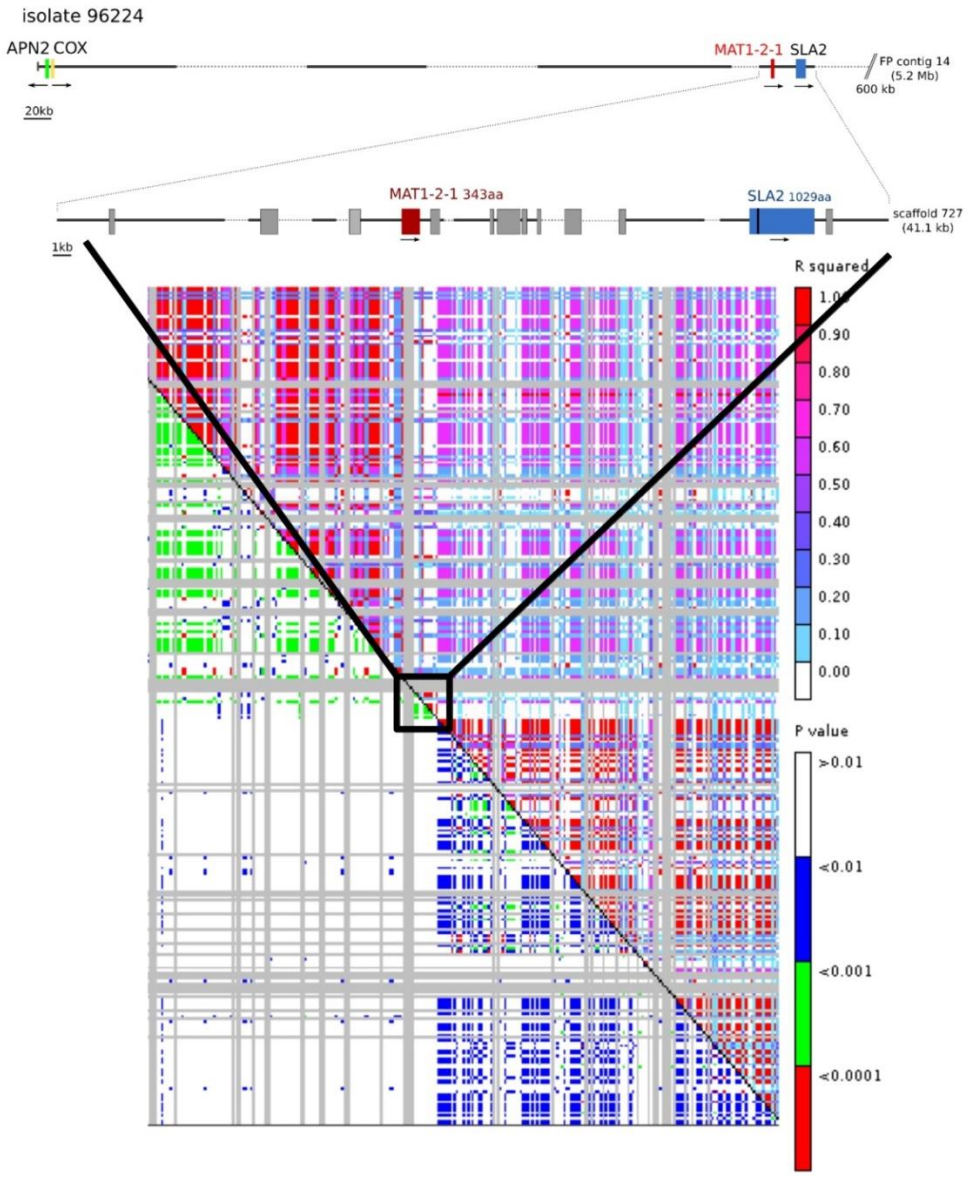
Supplementary Fig. 37. Phylogeographic analysis of *B.g. triticale* performed with MrBayes (Ronquist et al. 2012). Consensus tree of a partitioned analysis of the 1,152 genes for which all *B.g. triticale* isolates inherited the gene from one of the *B.g. tritici* parents. Posterior probabilities for individual clades are indicated if the values are smaller than one. The tree shows that *B.g. triticale* isolates and Swiss *B.g. tritici* isolates form a monophyletic group. This indicates that the genotype(s) of the *B.g. tritici* isolates which were involved in the original hybridization are more closely related to European than Israeli isolates and provides further evidence that Europe is the geographic origin of *B.g triticale* (Supplementary Note I).



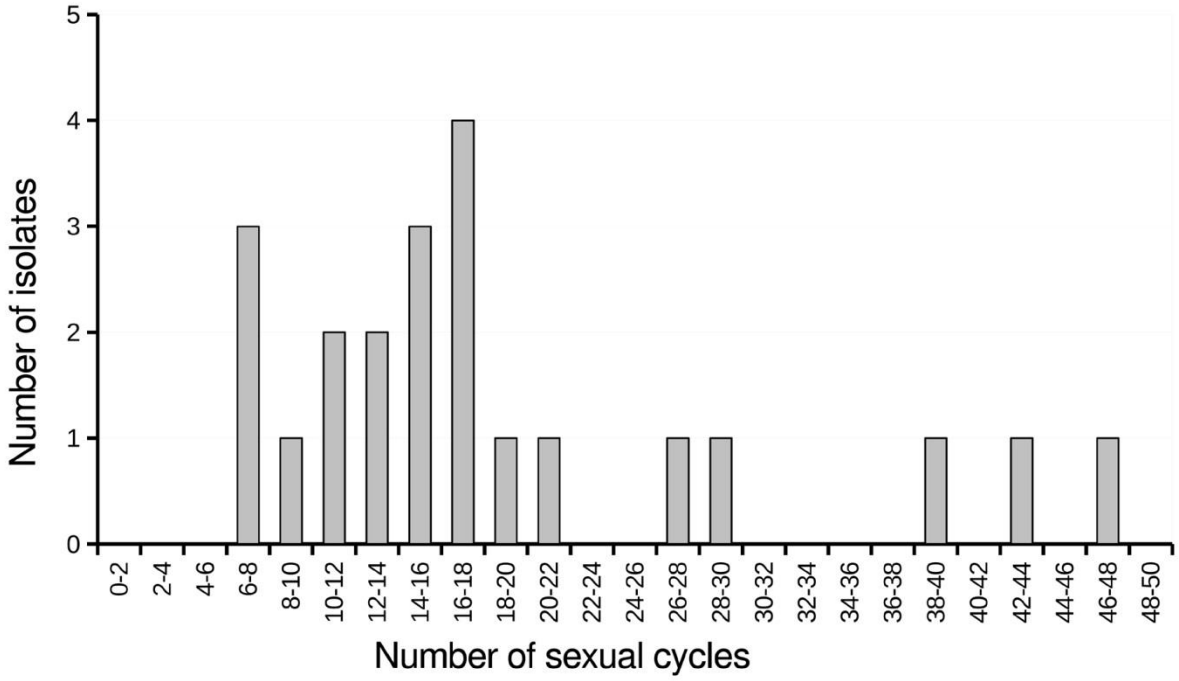
Supplementary Fig. 38. Example of haplotype counting for the gene *Bgt-2381*. This is one of the 64 single copy genes for which all *B.g. triticale* isolates cluster with *B.g. secalis*. In the tree *B.g. tritici* isolates are labelled in blue, *B.g. secalis* isolates in red and *B.g. triticale* isolates in black. In the alignment one can distinguish two different haplotypes (position indicated by an arrow). Haplotype A is represented by 3 *B.g. secalis* and 18 *B.g. triticale* isolates, while haplotype B is represented by two *B.g. secalis* and 4 *B.g. triticale* isolates. The two haplotypes are recognizable in the tree as well. Based on this type of analysis (extended to all genes with *B.g. secalis* genotype in all *B.g. triticale* isolates) we concluded that the minimum number of first hybridization event is two (Supplementary Note J).



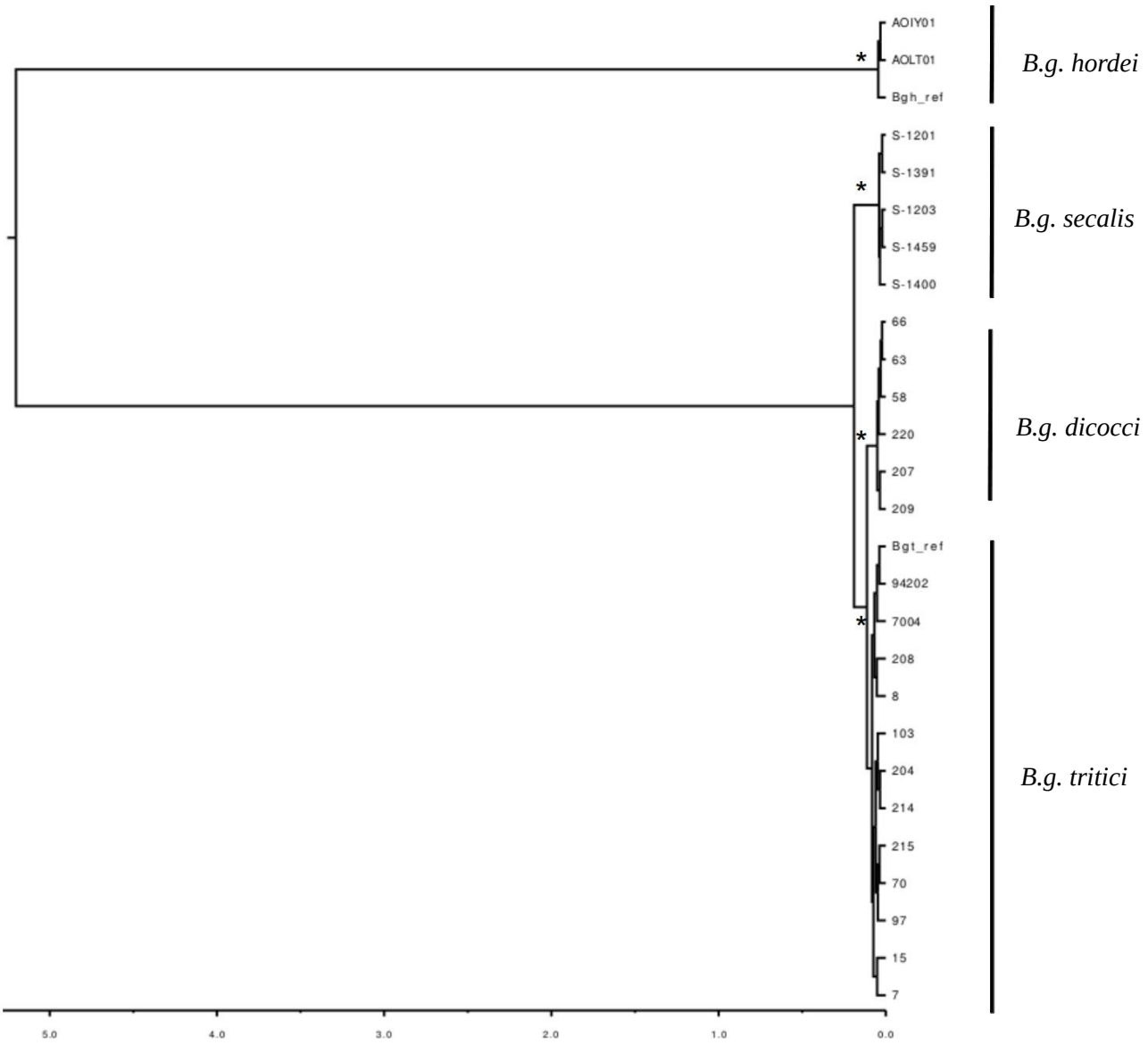
Supplementary Fig. 39. Model of hybridization followed by two back-crosses. The mating type MAT 1-2-1 was inherited from *B.g. secalis* in the first hybridization, a back-cross of a first generation hybrid with MAT 1-1-3 (inherited from *B.g. tritici*) with a *B.g. tritici* / MAT 1-2-1 isolate introduced the MAT 1-2-1 / *B.g. tritici* allele in *B.g. triticale* (Supplementary Note K).



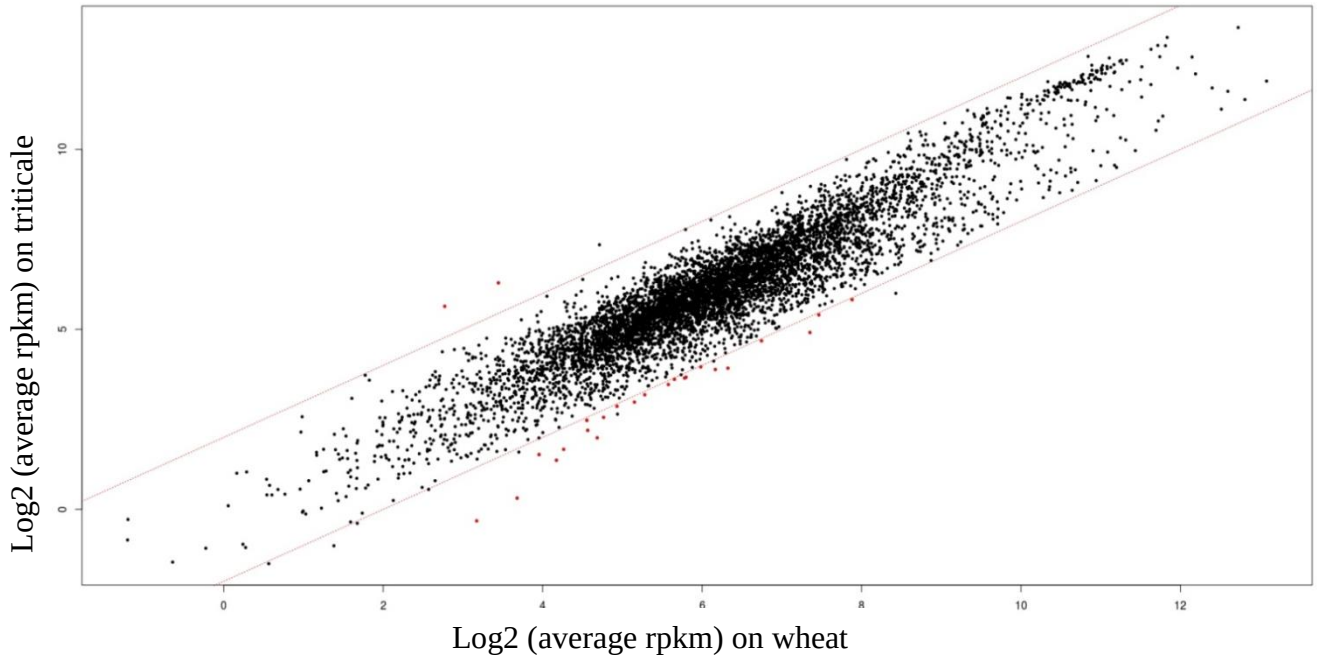
Supplementary Fig. 40. Analysis of Linkage Disequilibrium (LD) performed with TASSEL5 (Bradbury et al. 2007) at the mating type locus. The schematic representation of the mating type locus in the upper part of the figure is taken from Wicker et al. (2013). The mating type gene *Mat1-2-1* is located on scaffold 727 (on BAC contig 14) with another gene, *SLA2*. In the lower figure: on the right of the diagonal is represented r^2 for all pairwise SNPs in the first million bp of contig 14, high values of r^2 are represented in red shades, low values in blue shades (scale on the right); on the left of the diagonal the p-values of r^2 are shown. The color scale (on the right) goes from low values in red to higher values in white. The diagonal black line represents the LD of each SNP in the first million bp of contig 14 with itself. The mating type locus (highlighted in the black box) is not linked with neighboring scaffolds, the only linked gene is *SLA2* (Supplementary Note K).



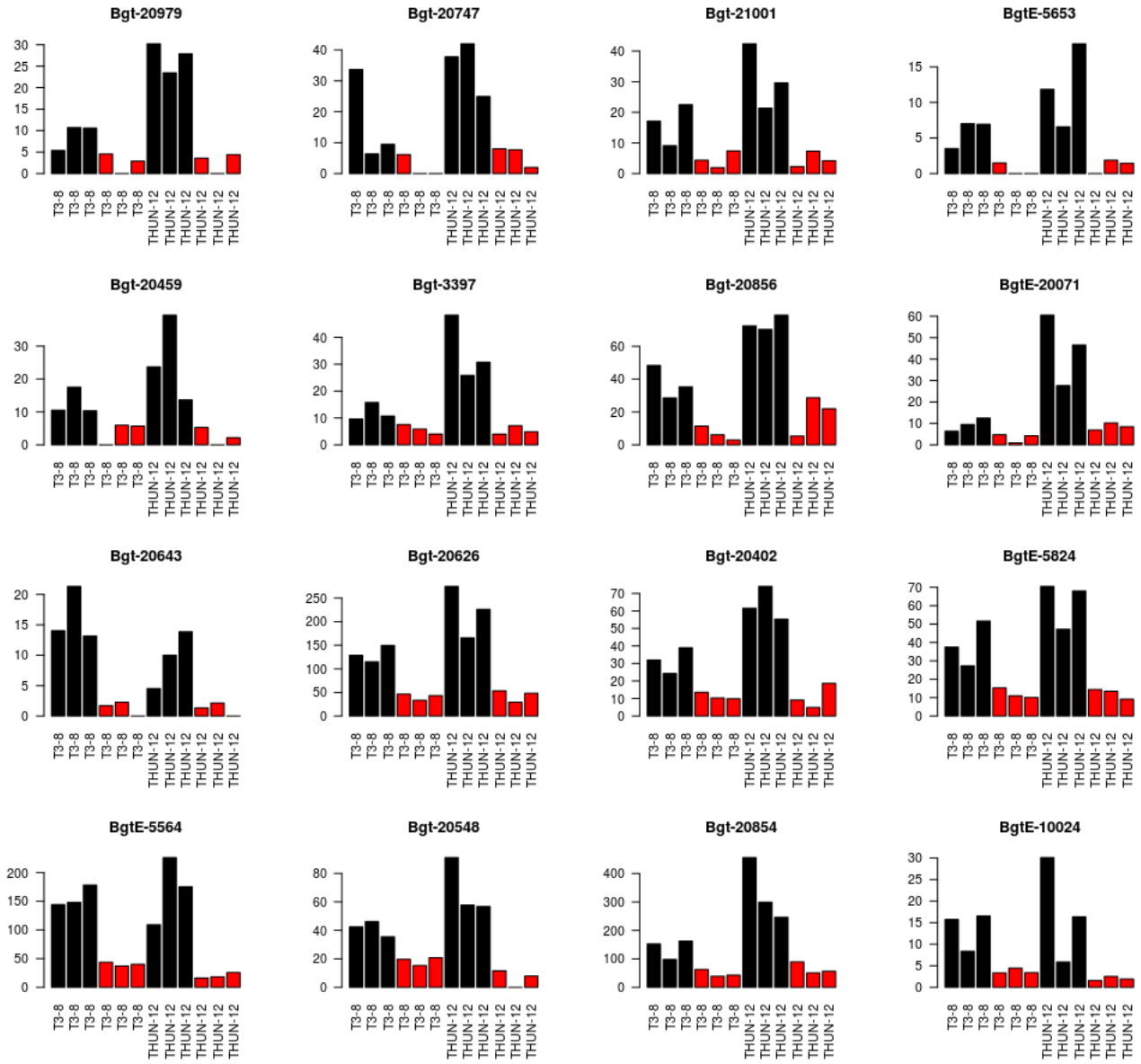
Supplementary Fig. 41. Number of sexual cycles from hybridization in *B.g. triticale*. These were estimated from the number of recombination break points between *B.g. tritici* and *secalis* genotypes in *B.g. triticale* (Supplementary Note M). Most of the isolates underwent between 6 and 22 sexual cycles confirming the young origin of *B.g. triticale*.



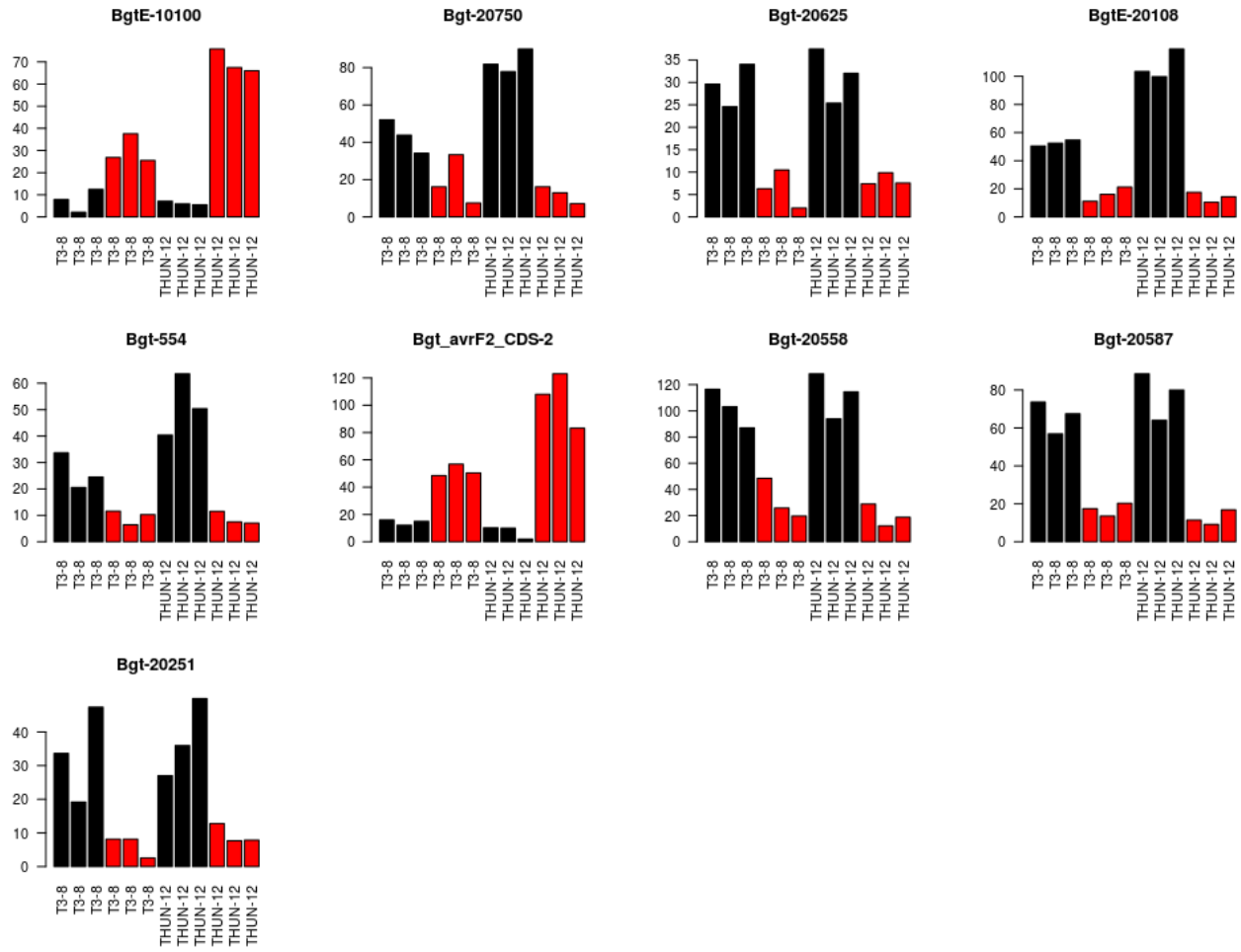
Supplementary Fig. 42. Bayesian consensus tree generated from an alignment of 206 orthologous genes from 27 powdery mildew isolates with MrBayes (Ronquist et al. 2012). The scale bar is in million years, clades supported by maximum posterior probability that correspond to *formae speciales* are indicated with asterisks (*). The tree was inferred using *Neurospora crassa* as outgroup, which is not shown for graphical reasons (Supplementary Note N).



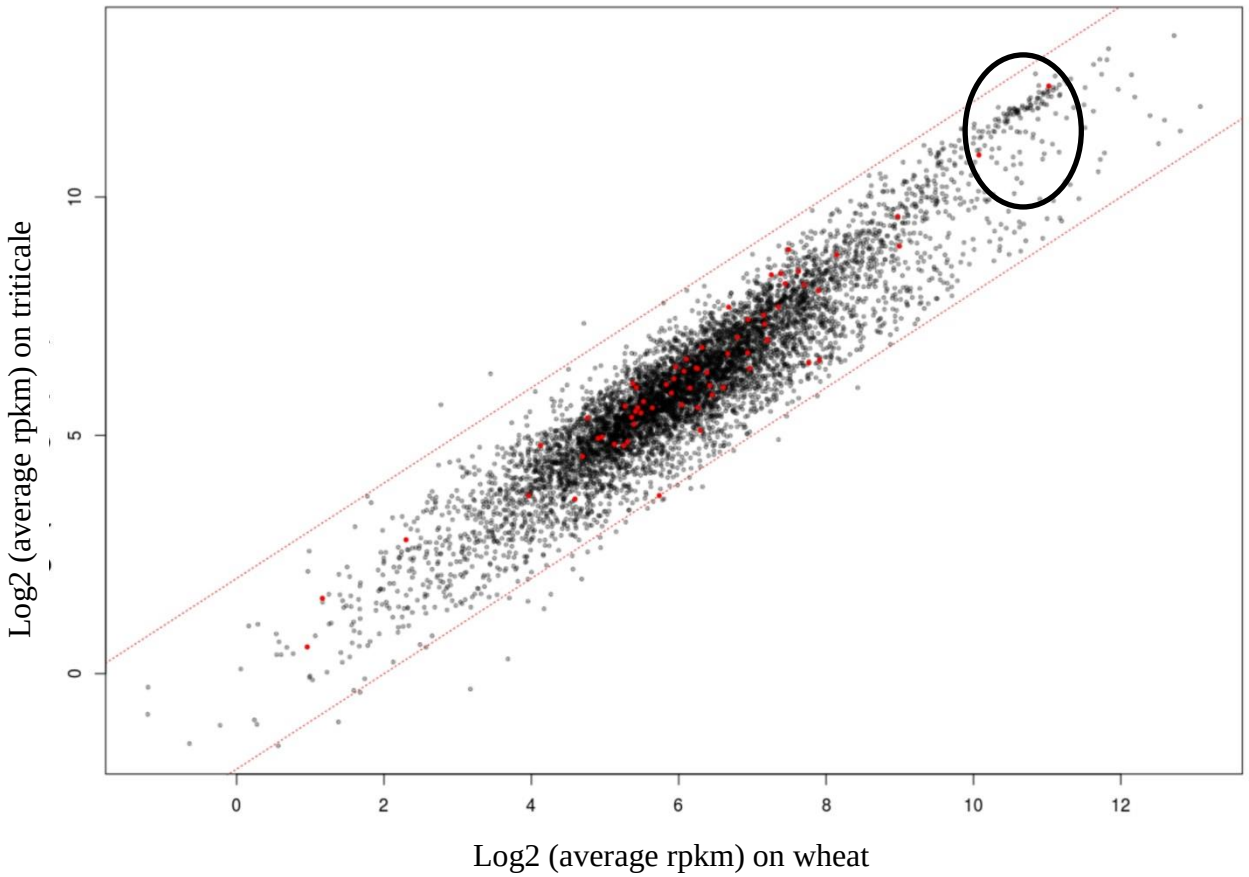
Supplementary Fig. 43. Plot of average changes in gene expression level of the two *B.g. triticale* isolates (T3-8 and THUN-12) in two different hosts (wheat and triticale). The values are shown in log2 of the number of reads per Kb per million mapped reads (rpkm). The two red lines represent the threshold of two log2 fold change in expression (equal to a 4-fold change). Note that the general expression is very similar in the two hosts and there are no genes that are highly differentially regulated. Only 25 genes (red dots) are significantly differently expressed in the two hosts (fdr corrected p-value < 0.001) with more than a 4-fold change. Most of these 25 genes are more highly expressed in wheat than in triticale (Supplementary Note P).



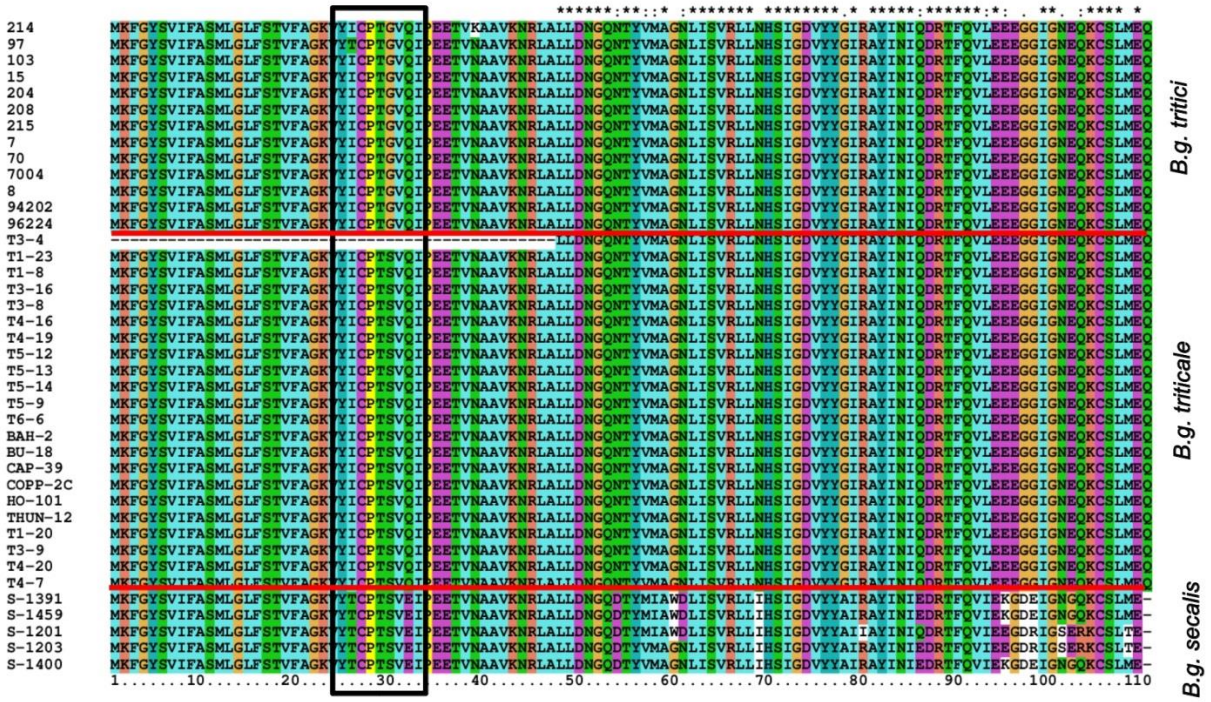
Supplementary Fig. 44a. Barplots of expression levels (number of reads per Kb per million mapped read, rpkm) of 16 of the 25 genes differentially expressed by *B.g. triticale* on the two different hosts (wheat and triticale) with an average fold change greater than four. We used two *B.g. triticale* isolates (THUN-12 and T3-8) on the wheat variety Chinese Spring (in black) and on the triticale variety Timbo (in red). Most of these genes have a low level of expression with rpkm values < 60, Median rpkm values on all genes are between 60 and 65 depending on the isolate / host combination (Supplementary Note P).



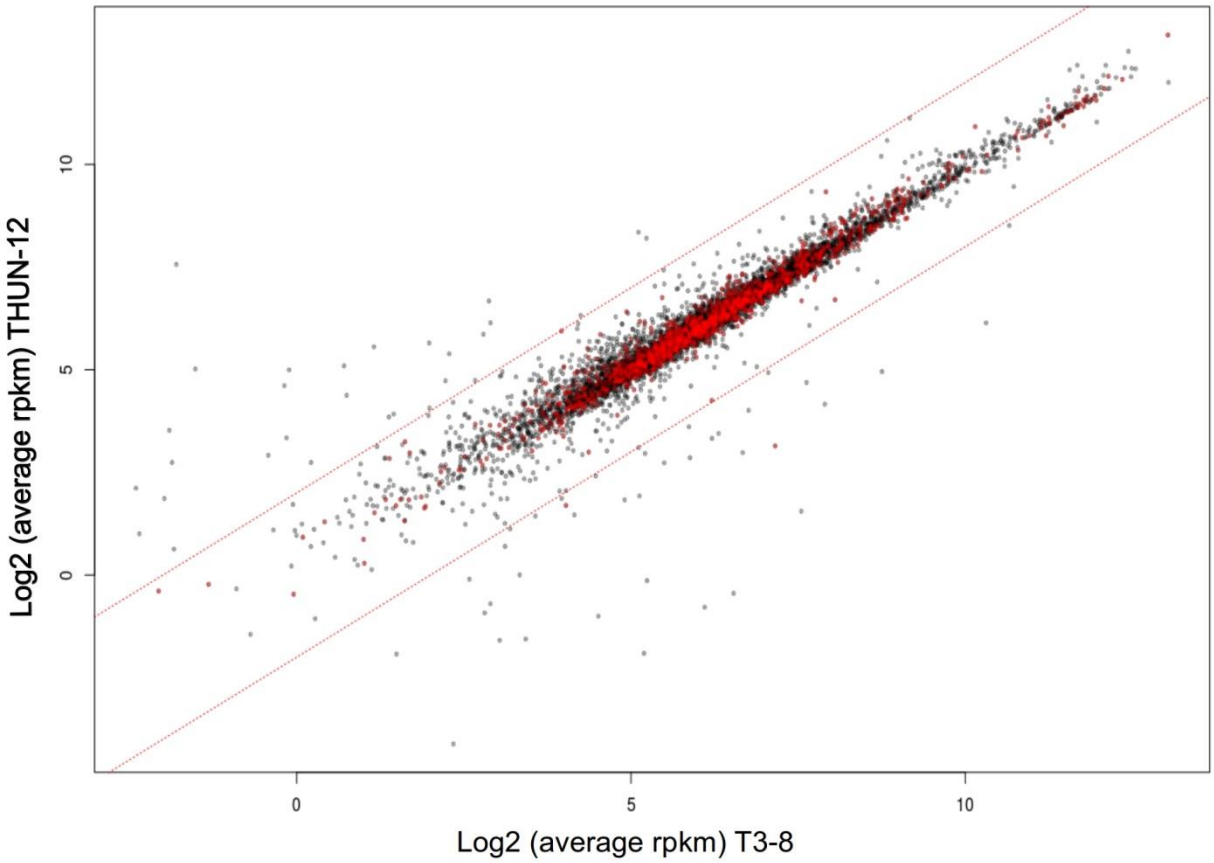
Supplementary Fig. 44b. Barplots of expression levels (number of reads per Kb per million mapped read, rpkm) of 9 of the 25 genes differentially expressed by *B.g. triticale* on the two different hosts (wheat and triticale) with an average fold change greater than four. We used two *B.g. triticale* isolates (THUN-12 and T3-8) on the wheat variety Chinese Spring (in black) and on the triticale variety Timbo (in red). Most of these genes have a low level of expression with rpkm values < 60, Median rpkm values on all genes are between 60 and 65 depending on the isolate / host combination (Supplementary Note P).



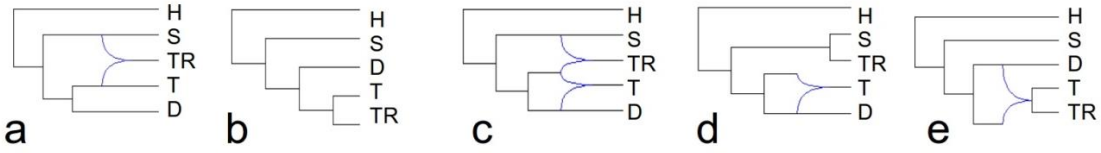
Supplementary Fig. 45. Plot of average changes in gene expression level of the two *B.g. triticale* isolates (T3-8 and THUN-12) in two different hosts (wheat and triticale). The data are the same as those shown in Supplementary Fig. 43. The values are shown in log2 of the number of reads per Kb per million mapped reads (rpkm). The two red lines represent the threshold of two log2 fold change in expression (equal to a 4-fold change). The 66 genes with a *B.g. secalis* genotype in all *B.g. triticale* isolates are shown in red. None of them has a log2 fold change greater than two and generally they are not differentially expressed in the two hosts (fdr corrected p-value < 0.001). Interestingly the two most highly expressed genes (in the black circle) among the 66 are putative effectors (Supplementary Note P).



Supplementary Fig. 46. Protein alignment of the putative effector *BgtAcSP- 31175. B.g. triticale* isolates carry a copy of the gene that has a *B.g. tritici* haplotype in all positions except one substitution where all isolates (except T3-4 that has a deletion of the first part of the gene) have the *B.g. secalis* genotype. This substitution causes an amino-acid mutation, from a glycine in *B.g. tritici* isolates to a serine in *B.g. secalis* and *B.g. triticale* isolates (in the black box). This gene could be the result of a double recombination between *B.g. secalis* and *tritici* genotypes in *B.g. triticale*. Alternatively the *B.g. tritici* parent that originated *B.g. triticale* had the same polymorphism but none of the sequenced isolate has it. Finally this pattern could also be the result of gene conversion.



Supplementary Fig. 47. Plot of average gene expression levels of the two *B.g.triticale* isolates T3-8 and THUN-12. The gene expression is computed as the average of the six replicates, in three replicates the isolates were grown on triticale, in the remaining three on wheat. The two red lines represent the threshold of two log2 fold change in expression (equal to 4-fold change). Note that there are more differentially expressed genes between isolates on the same host than in *B.g. triticale* grown on two different hosts (wheat and triticale) (Supplementary Figs. 43 and 45). The 1,304 single copy genes for which one of the two *B.g. triticale* isolates has a *B.g. tritici* genotype and the other a *B.g. secalis* genotype are represented in red. Only two of them are differentially expressed between the two isolates (fdr corrected p-value < 0.001) with more than two log2 fold change in expression. This indicates that variation in gene expression levels between *B.g. triticale* isolates does not depend on the parental origin of the allele (Supplementary Note P).



a: (H,((S,#H1),(D,(T,(TR)#H1))));
b: (H,(S,(D,(T,TR))));
c: (H,((S,#H2),((D,#H1),((T)#H1,(TR)#H2))));
d: (H,((S,TR),((D,#H1),((T)#H1))));
e: (H,(S,((D,#H1),((T,TR)#H1))));

H = *B.g. hordei*
S = *B.g. secalis*
D = *B.g. dicocci*
T = *B.g. tritici*
TR = *B.g. triticale*

Supplementary Fig. 48. Graphic (Dendroscope (Huson and Scornavacca 2012)) representation of the phylogenetic networks tested with PhyloNet (Than et al. 2008). In network a *B.g. triticale* is a hybrid between *B.g. secalis* and *B.g. tritici*. In network b there are no hybridization events. This tree corresponds to the tree in Supplementary Fig. 42 with the addition of *B.g. triticale* as a sister taxon of *B.g. tritici*. In network c *B.g. triticale* is a hybrid between *B.g. secalis* and *B.g. tritici* (as in network a), additionally also *B.g. tritici* is a hybrid between *B.g. dicocci* and an unknown lineage (as in networks d and e). Network c corresponds to the model in Fig. 3. In networks d and f *B.g. tritici* is a hybrid between *B.g. dicocci* and an unknown lineage. Alternatively *B.g. triticale* represent a sister taxon of *B.g. secalis* or of *B.g. tritici*. PhyloNet (Than et al. 2008) analysis showed network c to be the most likely (Supplementary Note Q).

II Supplementary Tables

Supplementary Table 1. List of *B. graminis* isolates sequenced with Illumina HiSeq.

Isolate name	Origin	Collected	References	*f. sp.*	Host species	Reads	Coverage	Number of SNPs	Mating type
7	Israel	1990	Eshed- Dinoor collection**	*B.g. tritici*	Bread wheat	32,508,056	17	125,510	MAT1-2-1
8	Israel	1990	Eshed- Dinoor collection**	*B.g. tritici*	Bread wheat	41,526,329	22	137,370	MAT1-2-1
15	Israel	1990	Eshed- Dinoor collection**	*B.g. tritici*	Bread wheat	49,943,003	27	151,694	MAT1-1-3
58	Israel	1990	Eshed- Dinoor collection**	*B.g. dicocci*	Durum wheat	35,140,177	18	178,123	MAT1-2-1
63	Israel	1990	Eshed- Dinoor collection**	*B.g. dicocci*	Durum wheat	47,363,151	22	195,493	MAT1-1-3
66	Israel	1990	Eshed- Dinoor collection**	*B.g. dicocci*	Durum wheat	45,767,674	26	205,147	MAT1-2-1
70	Israel	1990	Eshed- Dinoor collection**	*B.g. tritici*	Bread wheat	46,383,951	25	150,993	MAT1-2-1
97	Israel	1990	Eshed- Dinoor collection**	*B.g. tritici*	Bread wheat	46,276,460	23	144,688	MAT1-1-3
103	Israel	1990	Eshed- Dinoor collection**	*B.g. tritici*	Bread wheat	31,039,416	16	127,112	MAT1-1-3
204	Israel	2010	Eshed- Dinoor collection**	*B.g. tritici*	Bread wheat	55,533,513	31	150,864	MAT1-1-3
207	Israel	2010	Eshed- Dinoor collection**	*B.g. dicocci*	Durum wheat	36,380,111	19	187,023	MAT1-2-1
208	Israel	2010	Eshed- Dinoor collection**	*B.g. tritici*	Bread wheat	35,823,066	20	128,805	MAT1-1-3
209	Israel	2010	Eshed- Dinoor collection**	*B.g. dicocci*	Durum wheat	37,389,761	19	156,674	MAT1-1-3
215	Israel	2010	Eshed- Dinoor collection**	*B.g. tritici*	Bread wheat	41,681,063	23	147,781	MAT1-1-3
217	Israel	2010	Eshed- Dinoor collection**	*B.g. tritici*	Bread wheat	39,516,872	21	139,904	MAT1-1-3
220	Israel	2010	Eshed- Dinoor collection**	*B.g. dicocci*	Durum wheat	43,885,971	21	189,237	MAT1-2-1
7004	Switzerland	2007	Susanne Brunner	*B.g. tritici*	Bread wheat	42,197,291	22	116,501	MAT1-2-1
94202	Switzerland	1994	Wicker et al. 2013	*B.g. tritici*	Bread wheat	59,247,977	31	115,543	MAT1-1-3
BAH-2	Poland	2009-2010	Troch et al. 2012	*B.g. triticale*	Triticale	32,670,672	18	133,015	MAT1-2-1
BU-18	Belgium	2009-2010	Troch et al. 2012	*B.g. triticale*	Triticale	34,887,590	20	156,609	MAT1-2-1
CAP-39-A1	France	2009-2010	Troch et al. 2012	*B.g. triticale*	Triticale	40,842,225	22	146,136	MAT1-1-3
COPP-2C	France	2009-2010	Troch et al. 2012	*B.g. triticale*	Triticale	41,432,121	24	147,472	MAT1-2-1
HO-101	Belgium	2009-2010	Troch et al. 2012	*B.g. triticale*	Triticale	37,241,876	22	149,131	MAT1-2-1
S-1201	Switzerland	2013	Agroscope Nyon collection	*B.g. secalis*	Rye	40,815,971	21	331,818	MAT1-2-1
S-1203	Switzerland	2013	Agroscope Nyon collection	*B.g. secalis*	Rye	35,162,791	18	292,817	MAT1-1-3
S-1391	Switzerland	2013	Agroscope Nyon collection	*B.g. secalis*	Rye	38,433,116	21	326,019	MAT1-1-3
S-1400	Switzerland	2013	Agroscope Nyon collection	*B.g. secalis*	Rye	35,916,550	19	310,069	MAT1-2-1
S-1459	Switzerland	2013	Agroscope Nyon collection	*B.g. secalis*	Rye	40,762,910	22	332,450	MAT1-2-1
T1-20	Switzerland	2013	This study	*B.g. triticale*	Triticale	52,151,567	31	171,671	MAT1-2-1
T1-23	Switzerland	2013	This study	*B.g. triticale*	Triticale	43,202,785	24	155,702	MAT1-2-1
T1-8	Switzerland	2013	This study	*B.g. triticale*	Triticale	41,969,783	24	161,992	MAT1-2-1
T3-16	Switzerland	2013	This study	*B.g. triticale*	Triticale	37,462,675	22	149,479	MAT1-2-1
T3-4	Switzerland	2013	This study	*B.g. triticale*	Triticale	33,512,871	18	139,574	MAT1-1-3
T3-8	Switzerland	2013	This study	*B.g. triticale*	Triticale	43,669,157	25	156,499	MAT1-2-1
T3-9	Switzerland	2013	This study	*B.g. triticale*	Triticale	38,601,974	23	153,519	MAT1-2-1
T4-19	Switzerland	2013	This study	*B.g. triticale*	Triticale	37,718,984	21	157,750	MAT1-2-1
T4-20	Switzerland	2013	This study	*B.g. triticale*	Triticale	62,834,815	38	163,383	MAT1-2-1
T4-6	Switzerland	2013	This study	*B.g. triticale*	Triticale	38,920,466	23	144,351	MAT1-1-3
T4-7	Switzerland	2013	This study	*B.g. triticale*	Triticale	42,178,865	25	166,217	MAT1-1-3
T5-12	Switzerland	2013	This study	*B.g. triticale*	Triticale	38,758,888	21	145,964	MAT1-1-3
T5-13	Switzerland	2013	This study	*B.g. triticale*	Triticale	42,348,571	24	158,461	MAT1-1-3
T5-14	Switzerland	2013	This study	*B.g. triticale*	Triticale	46,861,574	27	159,926	MAT1-2-1
T5-9	Switzerland	2013	This study	*B.g. triticale*	Triticale	48,468,125	28	157,312	MAT1-2-1
T6-6	Switzerland	2013	This study	*B.g. triticale*	Triticale	41,178,865	23	145,121	MAT1-2-1
THUN-12	Poland	2009-2010	Troch et al. 2012	*B.g. triticale*	Triticale	45,874,763	27	164,253	MAT1-1-3
96224*	Switzerland	1996	Wicker et al. 2013	*B.g. tritici*	Bread wheat				MAT1-2-1
Mean						**41,810,764**	**23**	**171,670**	

**B.g. tritici* reference isolate (Wicker et al., 2013)[10]
**Hebrew University, Israel

Supplementary Table 2. Phenotyping of different *B. graminis* isolates on triticale, rye , tetraploid and hexaploid wheat.

	Species	TRITICALE						6n WHEAT			4n WHEAT		RYE		
	Cultivars	Timbo	Agroscope*	Triamant	Lamberto	Tridel	Bedretto	Kanzler	Chancellor	CS**	Langdon	Inbar	Matador	Sellino	Palazzo
F. sp.	Isolates														
B.g. secalis	S-1459	0	0	0	0	0	0	0	0	n.t.	n.t.	0	1	1	part
B.g. secalis	S-1201	0	0	0	0	0	0	0	0	n.t.	n.t.	0	1	1	1
B.g. secalis	S-1391	0	0	0	0	0	0	0	0	n.t.	n.t.	0	1	1	1
B.g. secalis	S-1203	0	0	0	0	0	0	0	0	n.t.	n.t.	0	1	1	part
B.g. secalis	S-1400	0	0	0	0	0	0	0	0	n.t.	n.t.	0	1	1	1
B.g. triticale	T4-19	1	1	1	1	1	1	1	1	1	n.t.	1	weak	0	0
B.g. triticale	T3-16	1	1	1	1	1	1	1	1	1	n.t.	1	0	0	0
B.g. triticale	T3-9	1	1	1	1	1	1	1	1	1	n.t.	1	weak	weak	0
B.g. triticale	T1-23	1	1	1	1	1	1	1	1	1	n.t.	1	0	weak	0
B.g. triticale	T1-8	1	1	1	1	1	1	1	1	0	n.t.	1	0	0	weak
B.g. triticale	T4-7	1	1	1	1	1	1	1	1	1	n.t.	1	weak	0	weak
B.g. triticale	T6-6	1	1	1	1	1	1	1	1	part	n.t.	1	0	0	0
B.g. triticale	T1-20	1	1	1	1	1	1	1	1	1	n.t.	1	0	0	0
B.g. triticale	T3-8	1	1	1	1	1	1	1	1	1	n.t.	1	0	0	0
B.g. triticale	T4-20	1	1	1	1	1	1	1	1	1	n.t.	1	0	0	0
B.g. triticale	T3-4	1	1	1	1	1	1	1	1	1	n.t.	1	weak	0	weak
B.g. triticale	T5-9	1	1	1	1	1	1	1	1	1	n.t.	1	0	weak	weak
B.g. triticale	T5-12	1	1	1	1	1	1	1	1	1	n.t.	1	0	0	0
B.g. triticale	T5-14	1	1	1	1	1	1	1	1	1	n.t.	n.t.	weak	0	weak
B.g. triticale	T4-6	1	1	1	1	1	1	1	1	1	n.t.	1	0	0	0
B.g. triticale	COPP-2C	1	1	1	1	1	1	1	1	1	n.t.	1	0	0	0
B.g. triticale	THUN-12	1	1	1	1	1	1	1	1	1	n.t.	1	0	0	0
B.g. triticale	BU-18	1	1	1	1	1	1	1	1	1	n.t.	1	weak	weak	0
B.g. triticale	CAP-39	1	1	1	1	1	1	1	1	1	n.t.	part	0	0	0
B.g. triticale	HO-101	1	1	1	1	1	1	1	1	1	n.t.	1	weak	0	0
B.g. triticale	BAH-2	1	part	1	1	n.t.	part	1	1	part	1	1	0	0	0
B.g. triticale	T5-13	part	1	1	1	1	1	1	1	1	n.t.	1	0	0	0
B.g. tritici	96224	0	0	0	0	n.t.	n.t.	1	1	1	1	1	0	0	n.t.
B.g. tritici	97	0	0	0	0	n.t.	n.t.	1	1	1	1	1	0	0	n.t.
B.g. tritici	94202	0	0	0	0	n.t.	n.t.	1	1	1	1	1	0	0	n.t.
B.g. tritici	208	0	0	0	0	n.t.	n.t.	1	1	1	1	1	0	0	n.t.
B.g. tritici	70	0	0	0	0	n.t.	n.t.	1	1	1	1	1	0	0	n.t.
B.g. tritici	204	0	0	0	0	n.t.	n.t.	1	1	1	1	1	0	0	n.t.
B.g. tritici	217	0	0	0	0	n.t.	n.t.	1	1	1	1	1	0	0	n.t.
B.g. tritici	103	0	0	0	0	n.t.	n.t.	1	1	1	1	1	0	0	n.t.
B.g. tritici	8	0	0	0	0	n.t.	n.t.	1	1	1	1	1	0	0	n.t.
B.g. tritici	15	0	0	0	0	n.t.	n.t.	1	1	1	1	1	0	0	n.t.
B.g. tritici	215	0	0	0	0	n.t.	n.t.	1	1	1	1	1	0	0	n.t.
B.g. tritici	7	n.t.	n.t.	n.t.	n.t.	n.t.	n.t.	1	n.t.	1	n.t.	n.t.	n.t.	n.t.	n.t.
B.g. tritici	7004	0	0	0	0	n.t.	n.t.	1	1	1	1	1	0	0	n.t.
B.g. dicocci	220	0	0	0	0	n.t.	n.t.	0	0	n.t.	1	1	0	0	n.t.
B.g. dicocci	63	0	0	0	0	n.t.	n.t.	0	0	n.t.	1	1	0	0	n.t.
B.g. dicocci	209	0	0	0	0	n.t.	n.t.	0	0	n.t.	1	1	0	0	n.t.
B.g. dicocci	58	n.t.	n.t.	n.t.	n.t.	n.t.	n.t.	0	n.t.	n.t.	n.t.	1	n.t.	n.t.	n.t.
B.g. dicocci	66	n.t.	n.t.	n.t.	n.t.	n.t.	n.t.	0	n.t.	n.t.	n.t.	1	n.t.	n.t.	n.t.
B.g. dicocci	207	n.t.	n.t.	n.t.	n.t.	n.t.	n.t.	0	n.t.	n.t.	n.t.	1	n.t.	n.t.	n.t.

1	Virulent (compatible interaction, pathogen can grow on host)
0	Avirulent
part	Partial host resistance presumably due to the presence of a race-specific resistance gene
weak	Limited growth of isolate observed
n.t.	Not tested

*Breeding line from Agroscope Nyon
**Chinese Spring

Supplementary Table 3. Contributions of parental *B.g. tritici* and *B.g. secalis* genomes to the genomes of *B.g. triticale* isolates.

Isolate	*B.g. tritici*[a]	*B.g. secalis*[b]	*B.g. triticale*[c]	W *B.g. tritici*[d]	L *B.g. tritici*[e]	W *B.g. secalis*[f]	L *B.g.secalis*[g]	Rec.[h]	L Rec.[i]	*B.g. secalis*[j]	Tot W[k]	Tot L[l]
BAH-2	146,639	25,608	27	13,369	58,025,633	1,351	4,106,647	1,479	9,546,446	6.61%	16,199	71,678,726
BU-18	135,551	36,694	29	12,637	55,419,401	2,313	7,784,423	1,250	8,476,748	12.32%	16,200	71,680,572
CAP-39	147,343	24,898	33	13,684	60,083,659	1,931	7,416,996	583	4,172,004	10.99%	16,198	71,672,659
COPP-2C	147,740	24,505	29	13,710	60,144,559	1,857	6,863,525	633	4,672,488	10.24%	16,200	71,680,572
HO-101	143,430	28,826	18	13,182	57,548,871	2,082	7,399,486	936	6,732,215	11.39%	16,200	71,680,572
T1-20	139,461	32,780	33	12,915	56,442,573	2,244	8,656,341	1,041	6,581,658	13.30%	16,200	71,680,572
T1-23	142,217	30,027	30	13,214	57,429,166	2,346	9,240,616	641	5,012,591	13.86%	16,201	71,682,373
T1-8	138,809	33,431	34	12,846	56,303,461	2,660	10,422,035	694	4,955,076	15.62%	16,200	71,680,572
T3-16	144,080	28,166	28	13,368	58,747,954	2,055	7,617,114	777	5,315,504	11.48%	16,200	71,680,572
T3-4	145,623	26,624	27	13,444	58,877,430	1,517	5,061,889	1,239	7,741,253	7.92%	16,200	71,680,572
T3-8	143,529	28,717	28	13,345	58,659,208	2,303	8,967,603	552	4,053,761	13.26%	16,200	71,680,572
T3-9	145,136	27,113	25	13,483	58,984,650	1,535	5,666,318	1,182	7,029,604	8.76%	16,200	71,680,572
T4-19	135,350	36,892	32	12,539	54,155,018	2,594	10,190,684	1,067	7,334,870	15.84%	16,200	71,680,572
T4-20	144,063	28,177	34	13,297	58,134,072	2,216	9,123,129	688	4,425,172	13.56%	16,201	71,682,373
T4-6	150,282	21,961	31	14,039	61,536,835	1,542	5,544,330	618	4,597,701	8.27%	16,199	71,678,866
T4-7	135,905	36,337	32	12,609	54,294,769	2,336	9,009,942	1,255	8,375,861	14.23%	16,200	71,680,572
T5-12	146,254	25,986	34	13,592	59,144,985	1,815	6,899,654	793	5,635,933	10.45%	16,200	71,680,572
T5-13	138,825	33,410	39	13,048	57,099,569	2,437	9,385,109	714	5,189,687	14.12%	16,199	71,674,365
T5-14	141,601	30,646	27	13,095	57,593,157	2,606	10,342,693	498	3,738,515	15.22%	16,199	71,674,365
T5-9	141,709	30,538	27	13,087	57,547,996	2,555	9,801,151	557	4,325,218	14.55%	16,199	71,674,365
T6-6	151,828	20,411	35	14,171	62,016,202	1,471	5,357,026	559	4,309,145	7.95%	16,201	71,682,373
THUN-12	138,708	33,539	27	12,783	56,071,169	2,903	11,727,920	513	3,875,276	17.30%	16,199	71,674,365

[a]Number of substitutions corresponding to *B.g. tritici* genotype
[b]Number of substitutions corresponding to *B.g. secalis* genotype
[c]Number of fixed substitutions with a third allele unique to *B.g. triticale*
[d]Number of windows with *B.g. tritici* genotype
[e]Cumulative length of windows with *B.g. tritici* genotype
[f]Number of windows with *B.g. secalis* genotype
[g]Cumulative length of windows with *B.g. secalis* genotype
[h]Number of recombinant windows
[i]Cumulative length of recombinant windows
[j]Proportion of *B.g. secalis* genotype (length *B.g. secalis* windows/length of total windows analysed)
[k]Total number of windows used for analysis
[l]Cumulative length of windows used for the analysis

Supplementary Table 4. Nucleotide diversity (π) and number of SNPs in *Blumeria graminis*.

Forma specialis	Total π	π in not CDS[1]	π in CDS[1]	Total number of SNPs	Intergenic SNPs	SNPs in exons (synonymous)	SNPs in intron
B. g. dicocci	0.0012	0.00128	0.0005	233,788	221,894	10,520 (4,556 / 43%)	1,374
B. g. secalis	0.00091	0.001	0.0003	156,292	150,375	5,232 (2,154 / 41%)	685
B. g. tritici	0.00169	0.00181	0.0007	522,733	496,493	23,018 (11,984 / 52%)	3,222
B. g. triticale	0.00145	0.00154	0.0006	386,049	365,390	18,007 (9,654/ 54%)	2,652

[1] Coding sequences

Supplementary Table 5. Estimates of divergence times of *B.g. tritici* and *B.g. hordei*.

Divergence	Reference	Basis for estimate
14,000 YA	Wyand and Brown, (2003)[13]	Host specificities of *ff. spp.*
4 .6 MYA	Inuma et al. (2007)[14]	ITS rDNA
6.3 MYA	Wicker et al., (2013)[10]	Synonymous substitutions in 5,258 genes
10 MYA	Oberhaensli et al., (2011)[15]	Orthologous transposable elements
11 MYA	Takamatsu and Matsuda, (2004)[16]	28S rDNA

Supplementary Table 6. List of the 66 genes with a *B.g. secalis* genotype in all *B.g. triticale* isolates. We report the conserved domain of each gene identified with CDD[17] (e-value < 0.001). Some genes have no identified conserved domain.

Gene	Conserved domain or putative function			
Bgt-1479	PNPOx_C	PNPOx/FlaRed_like		
Bgt-1492	STE2			
Bgt-154	Peptidase_C12			
Bgt-1597	EamA			
Bgt-1602	Thioredoxin_like			
Bgt-1644	TPP_enzymes			
Bgt-1646	Ribosomal_L7_L12			
Bgt-168	Sas10_Utp3			
Bgt-1692	GIY-YIG_SF			
Bgt-1696	P-loop_NTPase			
Bgt-1698	P-loop_NTPase			
Bgt-2088	Stc1			
Bgt-2289	PUA			
Bgt-2313	MMPL			
Bgt-2381	cyclophilin	RRM_SF		
Bgt-2458	Ntn_hydrolase			
Bgt-2459	Peptidase_C19	RHOD		
Bgt-255	ATP-grasp_4	Biotin_carb_C	CPSase_L_chain	Biotinyl_lipoyl_domains
Bgt-2688	TOM13			
Bgt-289				
Bgt-3062	SPOC	TFIIS_M	PHD_SF	
Bgt-3227	APC10-like			
Bgt-3241	WW	FF	PHA03283	
Bgt-3319	SLC5-6-like_sbd			
Bgt-3333	APP_MetAP	Creatinase_N		
Bgt-3338	P-loop_NTPase	Translation_factor_III	HBS1_N	DUF2967
Bgt-3339				
Bgt-3604	VHS_ENTH_ANTH	TrkH	PHD_SF	
Bgt-393	NOC3p	CBF		
Bgt-3967	RNA_lig_T4_1	tRNA_lig_kinase	tRNA_lig_CPD	
Bgt-4073	Kelch_1			
Bgt-4266				
Bgt-4518	Ribosomal_S16			
Bgt-4730	DUF2404			
Bgt-4889	YjeF_N			
Bgt-5109	chaperonin_like			
Bgt-5132	GAL4			
Bgt-5184	P-loop_NTPase	vWFA	ASF1_hist_chap	CobT
Bgt-5293				
Bgt-718	Thioredoxin_like			
Bgt-746				
Bgt-798	IDO			
Bgt-938				
Bgt-945	UBA_like_SF	SEP	UBQ	

Supplementary Table 6 (continuation).

Gene	Conserved domain or putative function
Bgt-973	NADB_Rossmann
BgtA-20329	SH3
BgtA-20582	Zip
BgtA-20629	P-loop_NTPase
BgtA-20637	Gpi16
BgtA-20738	AdoMet_MTases
BgtA-20840	WD40 F-box
BgtA-20892	Glycos_transf_3 Glycos_trans_3N
BgtA-20964	Esterase_lipase
BgtA-21035	
BgtA-21041	PHD_SF JmjN JmjC CTK3_C ARID zf-C5HC2 VHS_ENTH_ANTH
BgtA-21141	
BgtA-21152	F-box
BgtA-21389	DUF3449 zf-met SF3a60_bindingd
BgtAc-30654	
BgtAcSP-30775	Putative effector
BgtASP-20649	Putative effector
BgtE-20024	Putative effector
BgtE-5604	Putative effector LRAT
BgtE-5889	Putative effector
BgtE-5973	Putative effector

Supplementary Table 7. Results of PhyloNet with dataset 1, Lnlikelihood and three information criterion values for the five different networks for each gene set. The network with the highest information criterion value (in green) is considered the most likely. All information criteria gave the same results.
Composition of dataset 1: *B.g. hordei*: Bgh_ref; B.g. tritici: 96224, 97, 103; *B.g. secalis*: S-1201, S-1400; *B.g. triticale*: THUN-12, T3-8, COPP-2C; *B.g. dicocci*: 58, 66, 220;

Network (Number of parameters)		a (12)	b (8)	c(16)	d (13)	e (12)
LnLikelihood	Geneset1	-5,737.9	-5,840.9	-5,719.1	-5,896.6	-5,803.6
	Geneset2	-5,626.3	-5,705.7	-5,614.7	-5,771.9	-5,633.9
	Geneset3	-5,737.9	-5,840.9	-5,722.8	-5,896.6	-5,734.6
	Geneset4	-5,631.3	-5,705.6	-5,618.3	-5,757.7	-5,617.7
	Geneset5	-5,537.5	-5,615.8	-5,530.3	-5,667.3	-5,561.8
	Geneset6	-5,736.0	-5,805.5	-5,703.1	-5,901.1	-5,721.0
	Geneset7	-5,795.6	-5,838.8	-5,767.9	-5,886.2	-5,784.9
	Geneset8	-5,605.6	-5,680.4	-5,589.6	-5,745.1	-5,604.5
	Geneset9	-5,750.4	-5,840.5	-5,727.1	-5,915.9	-5,745.6
	Geneset10	-5,704.5	-5,787.6	-5,680.5	-5,828.4	-5,720.8
AIC	Geneset1	11,499.8	11,697.8	11,470.2	11,819.2	11,631.2
	Geneset2	11,276.6	11,427.4	11,261.4	11,682.8	11,291.8
	Geneset3	11,499.8	11,697.8	11,477.6	11,819.2	11,493.2
	Geneset4	11,286.6	11,427.2	11,268.6	11,541.4	11,259.4
	Geneset5	11,099.0	11,247.6	11,092.6	11,360.6	11,147.6
	Geneset6	11,496.0	11,627.0	11,438.2	11,828.2	11,466.0
	Geneset7	11,615.2	11,693.6	11,567.8	11,798.4	11,593.8
	Geneset8	11,235.2	11,376.8	11,211.2	11,516.2	11,233.0
	Geneset9	11,524.8	11,697.0	11,486.2	11,857.8	11,515.2
	Geneset10	11,433.0	11,591.2	11,393.0	11,682.8	11,465.6
AICc	Geneset1	11,487.8	11,689.3	11,455.0	11,806.4	11,619.2
	Geneset2	11,264.6	11,418.9	11,246.2	11,670.0	11,279.8
	Geneset3	11,487.8	11,689.3	11,462.4	11,806.4	11,481.2
	Geneset4	11,274.6	11,418.7	11,253.4	11,528.6	11,247.4
	Geneset5	11,087.0	11,239.1	11,077.4	11,347.8	11,135.6
	Geneset6	11,484.0	11,618.5	11,423.0	11,815.4	11,454.0
	Geneset7	11,603.2	11,685.1	11,552.6	11,785.6	11,581.8
	Geneset8	11,223.2	11,368.3	11,196.0	11,503.4	11,221.0
	Geneset9	11,512.8	11,688.5	11,471.0	11,845.0	11,503.2
	Geneset10	11,421.0	11,582.7	11,377.8	11,670.0	11,453.6
BIC	Geneset1	11,475.8	11,681.8	11,438.2	11,793.2	11,607.2
	Geneset2	11,252.6	11,411.4	11,229.4	11,656.8	11,267.8
	Geneset3	11,475.8	11,681.8	11,445.6	11,793.2	11,469.2
	Geneset4	11,262.6	11,411.2	11,236.6	11,515.4	11,235.4
	Geneset5	11,075.0	11,231.6	11,060.6	11,334.6	11,123.6
	Geneset6	11,472.0	11,611.0	11,406.2	11,802.2	11,442.0
	Geneset7	11,591.2	11,677.6	11,535.8	11,772.4	11,569.8
	Geneset8	11,211.2	11,360.8	11,179.2	11,490.2	11,209.0
	Geneset9	11,500.8	11,681.0	11,454.2	11,831.8	11,491.2
	Geneset10	11,409.0	11,575.2	11,361.0	11,656.8	11,441.6

Supplementary Table 8. Results of PhyloNet with dataset 2, Lnlikelihood and three information criterion values for the five different networks for each gene set. The network with the highest information criterion value (in green) is considered the most likely. All information criteria gave the same results.
Composition of dataset2: *B.g. hordei*: Bgh-ref; *B.g. tritici*: 94202, 8, 70; *B.g. secalis*: S-1203, S-1459; *B.g. triticale*: T1-23, T6-6, HO-101; *B.g. dicocci*: 63, 207, 209;

Network (Number of parameters)		a (12)	b (8)	c(16)	d (13)	e (12)
LnLikelihood	Geneset11	-5,664.7	-5,773.7	-5,644.9	-5,844.9	-5,696.8
	Geneset12	-5,695.0	-5,777.2	-5,662.6	-5,808.6	-5,697.2
	Geneset13	-5,709.5	-5,752.0	-5,839.3	-5,843.8	-5,690.4
	Geneset14	-5,734.9	-5,787.9	-5,684.6	-5,861.9	-5,701.1
	Geneset15	-5,759.0	-5,822.1	-5,743.1	-5,906.1	-5,784.1
	Geneset16	-5,731.9	-5,802.4	-5,715.3	-5,905.1	-5,749.1
	Geneset17	-5,912.0	-5,954.4	-5,888.7	-5,979.6	-5,901.0
	Geneset18	-5,855.5	-5,909.8	-5,850.8	-6,001.4	-5,825.6
	Geneset19	-5,699.0	-5,730.4	-5,698.7	-5,813.0	-5,663.9
	Geneset20	-5,625.0	-5,706.2	-5,599.2	-5,803.6	-5,623.3
AIC	Geneset11	11,353.4	11,563.4	11,321.8	11,715.8	11,417.6
	Geneset12	11,414.0	11,570.4	11,357.2	11,643.2	11,418.4
	Geneset13	11,443.0	11,520.0	11,710.6	11,713.6	11,404.8
	Geneset14	11,493.8	11,591.8	11,401.2	11,749.8	11,426.2
	Geneset15	11,542.0	11,660.2	11,518.2	11,838.2	11,592.2
	Geneset16	11,487.8	11,620.8	11,462.6	11,836.2	11,522.2
	Geneset17	11,848.0	11,924.8	11,809.4	11,985.2	11,826.0
	Geneset18	11,735.0	11,835.6	11,733.6	12,028.8	11,675.2
	Geneset19	11,422.0	11,476.8	11,429.4	11,652.0	11,351.8
	Geneset20	11,274.0	11,428.4	11,230.4	11,633.2	11,270.6
AICc	Geneset11	11,341.4	11,554.9	11,306.6	11,703.0	11,405.6
	Geneset12	11,402.0	11,561.9	11,342.0	11,630.4	11,406.4
	Geneset13	11,431.0	11,511.5	11,695.4	11,700.8	11,392.8
	Geneset14	11,481.8	11,583.3	11,386.0	11,737.0	11,414.2
	Geneset15	11,530.0	11,651.7	11,503.0	11,825.4	11,580.2
	Geneset16	11,475.8	11,612.3	11,447.4	11,823.4	11,510.2
	Geneset17	11,836.0	11,916.3	11,794.2	11,972.4	11,814.0
	Geneset18	11,723.0	11,827.1	11,718.4	12,016.0	11,663.2
	Geneset19	11,410.0	11,468.3	11,414.2	11,639.2	11,339.8
	Geneset20	11,262.0	11,419.9	11,215.2	11,620.4	11,258.6
BIC	Geneset11	11,329.4	11,547.4	11,289.8	11,689.8	11,393.6
	Geneset12	11,390.0	11,554.4	11,325.2	11,617.2	11,394.4
	Geneset13	11,419.0	11,504.0	11,678.6	11,687.6	11,380.8
	Geneset14	11,469.8	11,575.8	11,369.2	11,723.8	11,402.2
	Geneset15	11,518.0	11,644.2	11,486.2	11,812.2	11,568.2
	Geneset16	11,463.8	11,604.8	11,430.6	11,810.2	11,498.2
	Geneset17	11,824.0	11,908.8	11,777.4	11,959.2	11,802.0
	Geneset18	11,711.0	11,819.6	11,701.6	12,002.8	11,651.2
	Geneset19	11,398.0	11,460.8	11,397.4	11,626.0	11,327.8
	Geneset20	11,250.0	11,412.4	11,198.4	11,607.2	11,246.6

III Supplementary Notes

A. Taxonomic background, host specificity tests and crosses

B. Quantification of penetration efficiency in *B. graminis* at two days post infection

C. Results of re-sequencing of 45 powdery mildew isolates

D. Comparison between different SNP call pipelines

E. Population structure analysis

F. Genome wide nucleotide diversity patterns

G. Defining genomic windows for hybridization analyses

H. Quantification of the parental genome content in *B.g. triticale* and simulations of back crosses

I. Geographic origin of *B.g. triticale*

J. Estimation of the minimum number of isolates at the origin of *B.g. triticale*

K. Analysis of mating types

L. Background information on the life cycle of *B. graminis*

M. Estimates of numbers of sexual cycles in *B.g. triticale* since hybridization

N. Phylogenetic analysis and divergence time estimates of different *ff. spp*. in *Blumeria graminis*

O. Genes inherited from *B.g. secalis* in all *B.g. triticale* isolates

P. Transcriptome analysis

Q. Test of evolutionary models with PhyloNet

R. References

A. Taxonomic background, host specificity tests and crosses

Blumeria is a monospecific genus whose only species, *Blumeria graminis,* is divided in several *formae speciales* (*ff.spp.*). *Forma specialis* (*f.sp.*) is a taxonomic category that refers to a pathogen adapted to specific host species and that shows minimal (or no) morphological differences to its closest relative at the species level (Schulze-Lefert and Panstruga 2011). Eight *ff.spp.*have been described in *B. graminis*, each of them occurring on only one host genus (Troch et al. 2014). However several studies have shown that host specificity is not absolutely strict (Hardison 1943, Eshed and Wahl 1970, Troch et al. 2014). Troch and colleagues (2014) proposed to retain the category of *f.sp* only for *B. graminis* growing on cultivated cereals that show a stronger host specificity (*secalis, tritici, hordei* and *avenae*). The evolutionary origins of the different *ff. spp.* in *B. graminis* are subject of debate Panstruga and Spanu (2014): co-speciation between pathogen and host (Wicker et al. 2013), rapid diversification after the onset of agriculture (Wyand and Brown 2003, Troch et al. 2014) or the combined action of co-speciation and host jump or host range expansion (Inuma et al. 207) have been proposed as mechanisms for the origin of *ff.spp.*.

The different isolates tested for this study mostly showed a consistent host range in specificity tests, being specific for one, two or three species, depending on the *f.sp.* (Supplementary Table 2, Supplementary Fig. 2-23). However, there were a few exceptions to this general pattern: the intermediate phenotypes of *B.g. triticale* isolates T5-13 and BAH-2 on triticale cultivars Timbo, Agroscope and Bedretto and of *B.g. secalis* isolates S-1459 and S-1203 on rye cultivar Palazzo are likely the consequence of race-specific resistance genes in the cultivars Timbo, Agroscope, Bedretto and Palazzo (Supplementary Table 2). This interpretation is based on the fact that T5-13and BAH-2 are virulent on all the other triticale varieties and S-1459 and S-1203 are virulent on all the other rye varieties. Very limited growth of several *B.g. triticale* isolates was also observed on some rye cultivars (Supplementary Table 2, Supplementary Fig. 2-23), suggesting that *B.g. triticale* could grow not only on triticale and wheat but also to a low degree on rye. Limited growth and production of spores on non-host plants is relatively common in *B. graminis* (Troch et al. 2014). This could be important because it allows *ff. spp.*adapted to different hosts to still mate on non-host plants and therefore provide a possibility for hybridizations between *ff. spp*.

Prezygotic reproductive barriers in *B. graminis* have been shown to be weak. Indeed, the *ff. spp. secalis, tritici* and *agropyron* can produce fertile chasmotecia when crossed (Hiura 1978), but first generation hybrids have lower pathogenicity than the parents on both parental host species (Oku et al. 1986, Tosa 1989a and 1989b). To test whether isolates of the *f. sp. tritici* and *f. sp. triticale* can

mate, we performed a cross between the *B.g. tritici* reference isolates 96224 and the *B.g. triticale* isolate THUN-12 following the previously described protocols (Wicker et al. 2013). In that cross, the two isolates mated and produced several chasmothecia, indicating that *B. g. triticale* and *B.g. tritici* can potentially hybridize in nature. As negative control we grew the two parents alone under the same conditions and did not observe any chasmothecia formation.

B. Quantification of penetration efficiency in *B. graminis* at two days post infection

The macroscopic observation of infections on detached leaf segments after 10 days described in the host specificity assays in Supplementary Note A did not indicate any difference in pathogen growth between *B.g. tritici* and *B.g. triticale* isolates on wheat and between *B.g. triticale* isolates on wheat and triticale. However, the introgression of a considerable portion of the *B.g. secalis* genome in *B.g. triticale* could result in changes of growth parameters of *B.g. triticale* that can only be observed microscopically. We quantified penetration efficiency of one *B.g. tritici* (96224) and two *B.g. triticale* (THUN-12 and T3-8) isolates on two wheat varieties (Chinese Spring and Kanzler) and one variety of triticale (Timbo). The only macroscopic difference between these three isolates 10 days after infection was that 96224 did not grow on triticale. We observed infected leaves two days after infections after staining of fungal structures (Material and Methods). At this stage of infection it is possible to observe at the microscopic level single spores that germinated and attacked an epidermal cell. We estimated penetration efficiency as the proportion of infection attempts that resulted in the formation of a haustorium. For each plant/pathogen combination we counted at least 50 direct interactions on 4 different leaves (total > 200). Supplementary Fig. 24 shows that *B.g. triticale* has lower penetration efficiency than *B.g. tritici* on both varieties of wheat (chi-square test p-value < 0.001) compared to the wheat isolate. We hypothesize that this effect is due to introgression of *B.g. secalis* genome segments in *B.g. triticale* isolates. Moreover, we observed that *B.g. triticale* isolates have the same penetration efficiency on triticale and on the wheat variety Kanzler, while the penetration efficiency was lower on the wheat variety Chinese Spring. In conclusion, *B.g. triticale* has reduced fitness on both hosts, wheat and triticale, at least for penetration efficiency. However, this was macroscopically not observable after ten days.

C. Results of re-sequencing of 45 powdery mildew isolates

In total, 45 powdery mildew isolates were sequenced and compared to the reference genome

(Wicker et al. 2013). Highly repetitive regions of the genome are not included in the assembly of the reference genome of isolate 96224, while gene space is almost completely covered (Wicker et al. 2013). Thus, reads from re-sequenced isolates could only be mapped to the approximately 82 Mbp of assembled sequence of the reference genome. These mappings showed that all re-sequenced isolate genomes are highly similar to the genome of the reference isolate 96224, with sequences being on average 99.6%-99.8% identical. This high degree of sequence conservation allowed high-quality mapping and identification of polymorphic sites.

Illumina HiSeq sequencing of the 45 isolates resulted in between 31 and 62 million reads, corresponding to a sequence coverage between 16.2 and 37.9 (Supplementary Table 3). Mapping the reads to the 96224 reference isolated identified between 115,543 and 332,450 polymorphic sites, depending on the isolate (Supplementary Table 1).

D. Comparison between different SNP call pipelines

The goal of this analysis was to evaluate the quality of the SNP data used in the analysis where SNP calling was done with samtools[1] (Materials and Methods). We performed SNP calling on 12 isolates (isolates 217, 97 and 103 for *B.g. tritici*, S-1201, S-1459, S-1391 for *B.g secalis*, 209, 220 and 58 for *B.g. dicocci*, THUN-12, T4-19 and BAH-2 for *B.g. triticale*) using the two additional pipelines GATK 3.2-2 (DePristo et al. 2011) and FreeBayes 0.9.21-18 (Garrison and Marth 2012). We applied GATK mate pair fixing, duplicate removal and indel realignment and performed SNP discovery with the Unified Genotyper according to GATK Best Practices recommendations. FreeBayes was applied after duplicate removal.Variants were filtered with hard filtering parameters and compared using the VCFtools (Danecek et al. 2011). The results are summarized in Supplementary Fig. 25. If we consider the median value for each of the 12 isolates 91% of the SNP calls by samtools are called also by the other two pipelines.

The discordance between different SNP callers is well known and described in the literature (O'Rawe et al. 2013, Pirooznia et al. 2014, Pabinger et al. 2014). The caller that we used (samtools) produced only 9% of SNPs that are not called by both other used callers (GATK and FreeBayes). This is comparable to previous studies and normal for next generation sequences studies (O'Rawe et al. 2013, Pirooznia et al. 2014).

Additionally we performed a principal component analysis using the R package adegenet version 1.4-2 (Jombart and Ahmed 2011) on the SNPs called with the GATK pipeline (12 isolates) to check

if the result is similar to that of the principal component analysis shown in Fig. 2. Supplementary Fig. 26 shows that isolates from different *ff. spp.* cluster together in the same positions as in Fig. 2. Thus, we conclude that the technical variability resulting from different SNP call pipelines is small and does not affect the interpretation of biological differences in this study.

E. Population structure analysis

To determine population structure in the four *ff. spp.* we used the STRUCTURE 2.3.4 software (Pritchard et al. 2000). STRUCTURE assumes that loci are unlinked within populations. We randomly selected a subset of 10,000 of the available SNPs for the analysis, decreasing the SNP density to approximately one SNP every 8,000 bp. We performed a STRUCTURE analysis on all isolates used in this study and used the admixture model for 200,000 replications, 20,000 replications were used as burn-in. We ran this analysis for different values of K (from 1 to 6) and found that this dataset has a highest probability with K = 4 (Supplementary Fig. 27). With K = 2 *B.g. secalis* isolates are clustered in a group separated from all other isolates, with K = 3 *B.g. triticale* isolates formed an additional separate group, with K = 4 each *f. sp.* form a different group. Increasing K to 5 and 6 did not result in additional groups of isolates, at K=6 the probability of the data dropped, therefore we did not run additional analyses with higher values of K. Interestingly some *B.g. triticale* isolates are predicted to be admixed. They have SNPs inherited from two different ancestral populations, one of them in common with *B.g. tritici* isolates (Supplementary Fig. 27). We consider this as a further evidence of the hybridization of *B.g. triticale*. These results confirm the results of PCA and phylogeny that cluster the 4 *ff. spp.* in four different groups based on genomic data.

We then performed the same analysis individually for each *f. sp.* and found that for *B.g. tritici*, *B.g. dicocci* and *B.g. secalis*, the probability of the data is higher with K = 1 (we used K values from 1 to 4). Visual inspection of the barplots with K > 1 confirmed the absence of population structure: all isolates are composed by a mix of the ancestral populations with each population contributing for 1/Kth of the genotype of the isolate. This is considered as evidence of the absence of population structure by STRUCTURE's manual. We conclude that there is no obvious population structure in the sequenced isolates of these *ff. spp.*

For *B.g. triticale* the probability of the data increase with increasing values of K and reached a maximum with K = 5 (Supplementary Fig. 28). At K = 5 there are 5 groups of isolates that include:

T3-16, T3-4, T3-8 and T3-9 (group 1); T1-20 and T1-8 (group 2); T4-19 and T4-7 (group 3); T5-9 and T5-14 (group 4) and all the other isolates (group 5). The first 4 groups are composed of isolates that were collected the same day in the same location: group 1 in Reckenholz (CH), group 2 in Delley (CH), group 3 in Goumoens (CH), group 4 in Ellighausen (CH). Group 5 is composed of isolates from Switzerland, France, Belgium and Poland. With K = 6 group 5 was divided in two, the remaining Swiss isolates (T1-23, T4-20, T4-6, T6-6, T5-12 and T5-13) cluster together in a subgroup and isolates from the rest of Europe (BAH-2, BU-18, CAP-39, COPP-2C, HO-101and THUN-12) in another subgroup. Increasing K to 7 did not result in further separation of the described groups and with K = 8 the probability of the data dropped. We therefore did not run the analysis with higher values of K. We suspected that groups 1, 2, 3 and 4 identified with STRUCTURE's analysis are the result of a sampling bias because these isolates were collected at the same day in the same field, sometimes only few meters from each other. Therefore we hypothesized that these isolates are more closely related. To test this hypothesis we used GAPIT (Lipka et al. 2012) to construct a Kinship matrix (Supplementary Fig. 29) based on all SNPs and all isolates. The Kinship coefficient is an estimation of the probability that two isolates are identical (in one genomic position) by descent, thus that two alleles in two isolates are identical because they descend from an ancestral allele. The results of this analysis showed that isolates from group 1, 2, 3 and 4 identified with STRUCTURE have a higher probability to be identical by descent than any other pairwise combination of *B.g. triticale* isolates (Supplementary Fig. 29). This supports our hypothesis and shows that the population structure identified in *B.g. triticale* is probably due to a sampling bias. We then tested isolation by distance (IBD) on *B.g. triticale* and *B.g. tritici* (for *B.g. dicocci* and *secalis* we have only isolates from one country, Israel and Switzerland, respectively). The results (Supplementary Fig. 30) show that for *B.g. tritici* there is no correlation between genetic distance and geographic distance. *B.g. triticale* shows positive correlation, but this is the result of 9 outliers. These 9 outliers represent all possible pairwise combination of isolates of groups 1, 2, 3 and 4 defined by STRUCTURE and recognized to be related by the Kinship analysis. If we correct for this bias (excluding the pairwise comparisons inside groups 1, 2, 3 and 4), *B.g. triticale* does not show any isolation by distance. This can be expected from a species that originated recently and multiple times and moves by aerial dispersal of its spores (see below). For *B.g. tritici* we have samples from two populations (Switzerland and Israel), STRUCTURE analysis and IBD analysis did not detect divergence between these two populations. Powdery mildew spores can be transported by wind over several hundreds of km in a few days (Hermansen et al. 1978, Limpert et al. 1999, Brown and Hovømoller 2002). Hermansen et al. (1978) found that with favorable winds powdery mildew spores can be transported from England to the coast of Denmark traveling more

than 1,000 km in less than one day. Therefore, it is not surprising to find no distinct population structure and isolation by distance in *Blumeria graminis*, especially because between the two populations used in this study (Switzerland and Israel) wheat is grown virtually everywhere (Balkans, Turkey, Syria). In conclusion, our data show that *B.g. tritici* in Europe and the Middle East most likely belongs to a single panmictic population. However, the number of isolates used in this study is relatively small for population genetic studies and additional isolates covering a greater part of the geographic range of *B.g. tritici* are needed for definitive conclusions.

F. Genome wide nucleotide diversity patterns

To determine the intraspecific diversity of different *ff. spp.* and identify eventual patterns wer estimated nucleotide diversity (π) of different *ff. spp.* with the formula of Nei and Li (1979) with the software PopGenome (Pfeifer et al. 2014). In all *ff.spp.* diversity in intergenic regions was greater than in coding sequences, we also found that *B.g. tritici* and *B.g. triticale* have more synonymous substitutions (52% and 54%) compared to *B.g. dicocci* and *B.g. secalis* (43% and 41%) (Supplementary Table 4).

We additionally calculated π between pairs of isolates of the same *f. sp.* for all genome windows longer than 10 kb. *B.g. tritici, secalis* and *dicocci* showed a monomodal distribution of nucleotide diversity in genome windows (Fig. 2c and 2d). In contrast *B.g. triticale* showed a young hybrid diversity pattern with parts of the genomes with a high π and large portions of the genomes completely identical between isolates. This is due to the presence of the same haplotype (inherited from the same parental isolate in the initial cross or in one of the back-crosses) in different isolates (Fig. 2). Interestingly, this is reminiscent of the emergence of another fungal pathogen (*Zymoseptoria pseudotritici*), which was described in detail based on genomic data by Stukenbrock and colleagues (2012). *Z. pseudotritici* isolates also show portions of the genome with very low genetic diversity due to presence of the same haplotype inherited from one of the parents in several isolates. In particular in *Z. pseudotritici* the two parental genotypes have no diversity (only one haplotype from each of the parent was transmitted to the hybrid), this caused a complete loss of genetic diversity in about 30 % of the genome of *Z. pseudotritici* and it is due to the fact that only two individuals contributed to the hybridization. In *B.g. triticale* at least four individuals contributed to the first hybridization and the hybrid further backcrossed two times with a *B.g. tritici* isolate (Supplementary Notes H and J). Consequently *B.g. triticale* experienced a much reduced loss of diversity compared to *Z. pseudotritici*. Indeed the two parental genotypes in *B.g. triticale*

show a consistent part of the nucleotide diversity of the parents, 0.009 and 0.007 for the *B.g. tritici* and *B.g. secalis* genotypes respectively (π of *B.g. tritici* = 0.00169; π of *B.g. secalis* = 0.00091) (Supplementary Note G).

G. Defining genomic windows for hybridization analyses

We used genomic windows for several types of analysis that we describe in this chapter, the terminology of genomic windows is described in Supplementary Fig. 31. Windows of types A and D were defined as non-recombinant windows with the *B.g. tritici* genotype. We found between 12,539 and 14,171 such windows in the *B.g. triticale* isolates, accounting in average for 70.6% of the cumulative length of all windows. Type A and D windows common to all the *B.g. triticale* isolates (8,331; 35,731,610 bp; 44% of the genome) were used to compute the nucleotide diversity of the *B.g. tritici* genotypes in *B.g. triticale* (0.0009), the nucleotide diversity of *B.g. tritici* for the same genome segment is 0.0017.

Windows of type E are non-recombinant showing only the *B.g. secalis* genotype. We found between 1,351 and 2,903 such windows in the *B.g. triticale* isolates, accounting in average for 9.8% of the sequenced genome. Type E windows common to all the *B.g. triticale* isolates (51 windows; 125,331 bp) were used to compute the nucleotide diversity of the *B.g. secalis* genotypes in *B.g. triticale* (0.0007) and *B.g. secalis* (0.001) isolates. To estimate the proportion of the genome inherited from *B.g. secalis* in the *B.g triticale* isolates, we divided the cumulative length of non-recombinant windows with *B.g. secalis* genotypes (type E) by the cumulative length of all non-recombinant windows (types E + A + D).

Genomic windows that show segments of both *B.g. tritici* and *B.g. secalis* genotypes were defined as type B. These windows were not used for quantification of the portion of the genome inherited from the two parents, but they were divided into non-recombinant sub-windows (B_1 and B_2) and included in the analysis of genes originating from the *B.g. secalis* parent.

To test whether substitutions between *B.g. secalis* and *B.g. tritici* are randomly distributed in the genome we used type A, B and E windows with more than one substitution (82% of the genome). Genes were assigned to either of the two parents based on their genotype. Here, we used all non-recombinant (type E) and type B_2 sub-windows (the sub-windows with *B.g. secalis* genotypes).

Finally, to study nucleotide diversity distributions we used all windows longer than 10,000 bp (1,497 windows, 28% of the genome). Here, type C windows were also used if they were longer than 10,000 bp.

H. Quantification of the parental genome content in *B.g. triticale* and simulations of back crosses

To quantify the proportion of genome segments that were inherited from the two parents (*B.g. tritici* and *B.g. secalis*) in the *B.g. triticale* isolates, we first tested whether the polymorphic sites are randomly distributed in the genome. This would make the number of substitutions a direct proxy for the parental genome portions. We divided the *B.g. tritici* reference genome in to windows that do not contain sequence gaps and obtained 22,271 windows (average length 3,682 bp). A total of 13,760 windows (corresponding to 82% of the reference genome) contained two or more fixed polymorphisms (or substitutions) between *B.g. tritici* and *B.g. secalis* (Supplementary Fig. 31). We then calculated the genomic distances between substitutions. We obtained 156,072 distances and performed a Kolmogorov-Smirnov (KS) test between the cumulative probability distribution of the observed distances between substitutions and the cumulative probability distribution of simulated data from 156,073 SNPs in random positions on a chromosome that has the same size of the sum of all real distances (49,382,772 bp). The KS test rejected the null hypothesis that the observed distribution is equal to the simulated one with a p-value equal to $2.2*10^{-16}$, indicating that the observed substitutions are not randomly distributed in the genome and therefore cannot be directly used to estimate the proportion of parental genomes in the *B.g. triticale* isolates.

We then selected 16,201 genomic windows (cumulative length of 71,682,373 bp) that contain at least one substitution between *B.g. tritici* and *B.g secalis* (Supplementary Figure 31). For each *B.g. triticale* isolate we assigned these windows to two groups, one containing substitutions found in *B.g. tritici* genotypes and the other containing substitutions found in *B.g. secalis* genotypes. Using the cumulative lengths of these two groups of windows, we estimated the proportion of the *B.g. triticale* genomes inherited from *B.g tritici* and from *B.g. secalis*. We found that between 6.6 % and 17.3 % (median 12.8%) of the genome in *B.g. triticale* isolates is inherited from a *B.g. secalis* genotype, the rest from a *B.g. tritici* genotype (Supplementary Table 3, Supplementary Fig. 34-36). We hypothesized that this ratio of genome contributions is best explained by 2 back-crosses with *B.g. tritici* after hybridization and that this back-crossing process selected for the best adapted isolates for growth on triticale (Supplementary Fig. 48).

To evaluate the possibility that the hybrid pattern observed in *B.g. triticale* is due to artifacts we performed the following test: we hypothesized that *B.g. secalis* is a hybrid between *B.g. triticale* and *B.g. tritici* and performed the same test. If the test gives similar results to the original version in which we hypothesized *B.g. triticale* being a hybrid between *B.g. tritici* and *B.g. secalis* we should conclude that this test is not reliable. We identified substitutions between *B.g. triticale* and *B.g. tritici* and found 2,461 of them (compared to the 172,274 between *B.g. secalis* and *B.g. tritici*). These are positions inherited from *B.g. secalis* in all *B.g. triticale* isolates (indeed, we estimated this portion of the *B.g. triticale* genome to be ~ 0.18% (Supplementary Note G)). We then checked these positions in *B.g. secalis* isolates and found that more than 90% of these positions in *B.g. secalis* have a *B.g. triticale* genotype, while the remaining less than 10%, between 171 and 240 SNPs (depending on the isolate), have a *B.g. tritici* genotype. In this case these 171 to 240 SNPs represent polymorphisms that were present in the parental isolates of *B.g. triticale* (*tritici* or *secalis*) but not in the set of isolates that we sequenced. This is a surprisingly small number, given the relatively small sample size of *B.g. tritici* and *B.g. secalis* used for complete sequencing. We conclude that applying this test to species that are not hybrid does not result in a hybrid pattern. Moreover the obtained results can be fully explained by the hypothesis of *B.g. triticale* being a hybrid.

Simulations were performed to test the hypothesis of a hybridization followed by back-crossing. They were based on the observation that the proportion of *B.g. secalis* genome in *B.g. triticale* isolates is normally distributed (the Shapiro-Wilk normality test couldn't reject the null hypothesis of normality, p-value = 0.55) with a mean of 0.122 and standard deviation of 0.03. We assumed that when two isolates mate, their offspring inherit a certain proportion of the genome from the two parents that is normally distributed with a mean of 0.5. We performed the simulations with different values for the standard deviation (0.015, 0.03, 0.05). We simulated the first hybridization by randomly extracting a value (a) from a normal distribution with a mean of 0.5, the first back-cross was simulated by randomly extracting a value (b) from a normal distribution with mean of (a/2) and the second back-cross by randomly extracting a value (c) from a normal distribution with mean of (b/2). The value c represents the simulated proportion of the *B.g. secalis* genome in one single *B.g. triticale* isolate. We estimated the distribution of c with 10,000 independent replications. As expected, the obtained distribution was still normally distributed with a mean of 0.125 (Supplementary Fig. 34). Moreover, a KS test did not reject the null hypothesis that the observed and simulated data show the same distribution. The simulation above is simplistic and does not take into consideration mating between hybrid isolates. To study a possible effect of the latter we simulated 100 independent replications of a Fisher-Wright population (n=10,000) and let hybrid

isolates mate between themselves for 100 generations (with a standard deviation of the proportion of genome transmitted to the offspring of 0.02). The starting values of the proportion of *B.g. secalis* genome in hybrid isolates was simulated with the process described above. Mating between hybrid isolates increased the standard deviation of the distribution. This effect is very strong for the first generations but gets rapidly smaller (e.g. the difference between generation 1 and generation 2 is greater than the difference between generation 2 and generation 100, see Supplementary Fig. 35). Again, the observed data fit the simulated distribution at generation 1 and also after 100 generations (KS test p-value = 0.2568). We were then interested in studying the effect of constant, moderate gene flow from *B.g. tritici* to *B.g. triticale.* To do this we used the simulation above but at each generation 10 isolates (0.1% of the population) were assumed to be the result of a cross between a *B.g. triticale* isolate and a *B.g. tritici* isolate. As expected, the effect of this process is a shift in the distribution of the *B.g. secalis* proportion of genome in *B.g. triticale* to the left towards a lower proportion of *B.g. secalis* genome (Supplementary Fig. 36). With this moderate level of gene flow and this population size, the shift is slow and after 100 generations the KS test cannot reject the null hypothesis that the observed and simulated data are of the same distribution (p-value = 0.8964). In conclusion, the results of these simulations show that the observed proportion of *B.g. secalis* genome in *B.g. triticale* isolates is consistent with the hypothesis of one hybridization event that is followed by two back-crosses. They also show that a moderate gene flow from *B.g. tritici* (and probably also from *B.g. secalis*, not simulated) to *B.g. triticale* cannot be excluded. With this analysis we cannot exclude that other more complex scenarios would fit the data, however we show that the relatively simple model which includes one hybridization and two back-crosses fully explains the observed data.

I. Geographic origin of *B.g. triticale*

Powdery mildew disease on triticale was reported for the first time in Belgium, France and Poland (Walker 2011). Moreover to our knowledge this disease is so far confined to Europe and it is therefore likely that *B.g. triticale* originated in Europe. To test this hypothesis we performed a phylogeographic analysis. We inferred phylogeny of 4,556 single copy genes (Materials and Methods) and identified 3,206 gene trees in which *B.g. tritici* and *B.g. secalis* cluster monophyletically. The remaining trees do not have a phylogenetic signal and were thus not used for the analysis. This high proportion of trees without phylogenetic signal is indicative of a high level of incomplete lineage sorting that is characteristic of young species (Rosenberg and Nordborg 2002,

Maddison and Knowles 2006). For 1,152 of the trees with a phylogenetic signal all *B.g. triticale* isolates cluster with *B.g. tritici* isolates, indicating that the genes were inherited from the *B.g. tritici* parental isolates. We used these genes in a partitioned phylogenetic analysis with MrBayes 3.2.2 (Ronquist et al. 2012). We ran two independent replications of 10,000,000 generations. Variation of substitution rates across sites was modeled with a discretized (4 categories) gamma (Γ) distribution (Yang 1993 and 1994). The chains were let free to sample all the models of the GTR model family using a reversible jump Monte Carlo Markov Chain (Huelsenbeck et al. 2004). The results of this analysis (Supplementary Fig. 37) shows that all *B.g. triticale* isolates are more closely related to Swiss *B.g. tritici* isolates than to Israeli isolates. This finding further supports the hypothesis of a European origin of *B.g. triticale*. Moreover, *B.g. triticale* isolates collected from different European countries do not cluster in monophyletic groups, therefore is not possible to further narrow down the geographic area where *B.g. triticale* originated. It is known that powdery mildew spores can be transported over long distances by winds (Hermansen et al. 1978, Limpert et al. 1999, Brown and Hovømoller 2002). Thus, there is a high potential for gene flow between different populations in Europe which would obscure the actual site of origin. This is reflected by the absence of differentiation between European populations of *B.g. triticale* (Supplementary Note E).

J. Estimation of the minimum number of isolates at the origin of *B.g. triticale*

Given the hybrid origin of *B.g. triticale* we wanted to know if the hybridization occurred only once or several times. Interestingly, sequence segments in the genome of *B.g. triticale* isolates that have been inherited from the *B.g. secalis* parent show some polymorphism between isolates (Supplementary Notes F). This indicates that more than one *B.g. secalis* isolate contributed to the *B.g. triticale* hybrid and therefore the initial hybridization occurred more than once independently. To estimate more precisely the minimum number of isolates that have been involved we used a haplotype counting method that has been used in similar analyses (Shimizu-Inatsugi et al. 2009). Basically this method counts the number of different haplotypes inherited from one parent in the hybrid species, and this number represents the minimum number of isolates involved in the hybridization (Supplementary Fig. 39). We used the 64 single copy genes for which all *B.g. triticale* isolates cluster together with *B.g. secalis* (Supplementary Note O). On this set of genes we counted the number of different haplotypes present in *B.g. triticale*. We found that *B.g. triticale* isolates have completely identical sequences for 56 of these genes, 6 genes show two different haplotypes (observed also in *B.g. secalis*) and two genes show three haplotypes. The third haplotype is in both

cases represented by only one isolate, in one case it differs by a single SNP and in the other it is actually a chimera between the other two haplotypes. We conclude that in both cases the generation of the third haplotype could have occurred after the hybridization, through a single mutation or a recombination event between two *B.g. triticale* isolates.

We then further selected the 100 single copy genes that show the highest nucleotide diversity in *B.g. secalis* (to maximize the chance of having different haplotypes in the parent). Among them we identified 42 trees with a clear phylogenetic signal (defined as all *B.g. secalis* isolates grouped together in a monophyletic group), whereas all other trees were completely or almost completely unresolved with bootstrap values close to zero for most branches. The low amount of phylogenetic information is indicative of the young age of the two species and indicates a considerable amount of incomplete lineage sorting (Rosenberg and Nordborg 2002, Maddison and Knowles 2006). In 22 of the 42 trees with a phylogenetic signal, all *B.g. triticale* sequences cluster with *B.g. tritici*, while for the remaining 20 we identified *B.g. triticale* sequences that cluster with *B.g. secalis*. Of these 20, a total of 13 genes showed only one haplotype and 7 genes showed 2 haplotypes. Again we conclude that the minimum number of *B.g. secalis* isolates that were in involved in the evolution of *B.g. triticale* is two.

K. Analysis of mating types

As most fungi, *B. graminis* has two mating types, MAT1-2-1 and MAT1-1-3, and only isolates with opposite mating type can mate. The sequences of the two mating type loci and the associated SLA2 gene are known from previous sequence analyses (Wicker et al. 2013). This allowed the isolation of all mating type loci sequences from the genome assemblies of all the isolates studied here.

We found that two *B.g. secalis* isolates, 6 *B.g. tritici* and 16 *B.g. triticale* have the MAT1-2-1 mating type, while 3 *B.g. secalis*, 7 *B.g. tritici* and 6 *B.g. triticale* have the MAT1-1-3 type (Supplementary Table 1). The MAT1-2-1 + SLA2 genes have 7 fixed SNPs that distinguish all *B.g. tritici* from all *B.g. secalis* isolates. Three of the 16 *B.g. triticale* isolates that have the MAT1-2-1 mating type have the genotype of *B.g. secalis* in all 7 SNP position, the other 13 carry the *B.g. tritici* haplotype.

The MAT1-1-3 + SLA2 genes have three fixed SNPs that distinguish *B.g. tritici* and *B.g. secalis* genotypes. Here, all six *B.g. triticale* isolates contain the *B.g. tritici* genotype. This finding indicates that an initial hybridization probably involved only one combination of mating type: one or more

MAT1-1-3 *B.g. tritici* isolates that mated with one or more MAT1-2-1 *B.g. secalis* isolates. The MAT1-2-1 mating type with a *B.g. tritici* genotype was acquired by *B.g. triticale* with one of the back-crosses (Supplementary Note H, Supplementary Fig. 48). It is also possible that a MAT1-2-1 *B.g. tritici* / MAT1-1-3 *B.g. secalis* combination of isolates was involved but the *MAT1-1-3 B.g. secalis* allele was simply not present in our sample.

The mating-type loci in *B. graminis* evolve at a slower rate than in other fungal pathogens: We only found 1 and 3 amino acid changes between *B.g. secalis* and *B.g. tritici* which we estimate to have diverged 150,000 -250,000 years ago (Supplementary Note N). This is in contrast to the identified 5 and 20 amino acid changes which occurred during 10,000 years of evolution in *Z. pseudotritici* (Stukenbrock et al. 2011). High polymorphism between species at the mating type locus could be a barrier to gene flow. We observed that *B.g. tritici* and *B.g. secalis* are not very diverse at the mating type locus compared with other fungal pathogens (Stukenbrock et al. 2011). However, genome wide population genetic analysis, principal component analysis (PCA) and phylogenetic methods show clustering of isolates belonging to the same *f. sp.,* indicating the presence of strong barriers to gene flow (Supplementary Notes N and E). We conclude that postzygotic reproductive barriers are probably stronger than prezygotic ones in *B. graminis*.

We were interested in analyzing the pattern of linkage disequilibrium (LD) around the mating type locus in *B.g. triticale*. The absence of the *MAT1-1-3* mating type with a *B.g. secalis* genotype could be due to some gene linked to the *MAT1-1-3* locus for which *B.g. triticale* needs the *B.g. tritici* genotype to successfully conclude its life cycle. We therefore performed a LD analysis of the locus with TASSEL5(Bradbury et al. 2007). The result shows that the scaffold including the mating type locus is not linked with nearby scaffolds (Supplementary Fig. 40). This is probably the result of recombination hotspots around the mating type locus. Indeed two genetic markers, one on the mating locus and the other on the extremity of contig 14 are physically separated by only about 500 Kb but have a genetic distance of 14.9 centimorgan (cM) on the genetic map of *B.g. tritici* (Bourras et al. 2015).

L. Background information on the life cycle of *B. graminis*

B. graminis has a sexual and an asexual life cycle. The asexual cycle begins with a haploid conidiospore landing on the leaf and penetrating the epidermal plant cell wall after formation of an appressorium (Zhang et al. 2005). Inside the plant cell, the fungus forms a highly specialized

structure (haustorium) which is surrounded by the plasma membrane of the plant cells. This close association of haustorial surface and plant plasma membrane enables the assimilation of nutrients from the plant and probably promotes the transfer of fungal components into the plant cell. After that, the fungus grows secondary hyphae and produces new conidiospores which are then further distributed by wind. In nature the sexual cycle is triggered by dry weather at the end of summer, when hyphae of opposite mating types fuse and start a short diploid phase. Ascospores (sexual spores) are produced in fruiting bodies called chasmothecia. The mating type of powdery mildew is determined by a single genetic locus, the MAT1 locus (Coppin et al. 1997) (Supplementary Note K). Chasmothecia can remain dormant during rough weather conditions, allowing the fungus to over-winter or survive long periods of drought. It is known that chasmothecia can be stored for years under dry condition at room temperature or at 4° C, although capacity to eject ascospores decreases with time, while conidiospores are not viable few days after their detachment from the conidiophore[10]. However, asexual conidiospores are also known to survive winter on not harvested plant, and winter wheat seedlings ("green bridges") (Liu et al. 2012). Asexual fusion of hyphae was never observed in *B. graminis* and sexual reproduction has been described as the only opportunity of recombination between different powdery mildew strains Glawe 2008).

M. Estimates of numbers of sexual cycles in *B.g. triticale* since hybridization

Powdery mildew usually undergoes a sexual cycle in late summer. Because of the proposed recent origin of the hybrid *B.g. triticale*, only a limited number of sexual cycles could have occurred. Thus, we estimated the number of sexual generations after hybridization for each *B.g. triticale* isolate analyzed in this study. For this analysis we assumed that *B.g. triticale* originated through one hybridization and two back crosses (Supplementary Note H). We estimated the number of sexual generations after hybridization for each *B.g. triticale* isolate using the formula modified from Stukenbrock et al. (2012) (Materials and Methods).

Interestingly, we obtained estimates that range from 7 to 47 sexual cycles for the individual isolates (Supplementary Fig. 41). These values are consistent with our hypothesis of a recent origin of *B.g. triticale* approximately within the past 20-30 years. It is possible that some of the outliers (i.e. high values) are caused by gene conversion events that were not recognized with the criteria used here. Alternatively they could be the descendant of an older hybridization that occurred about 50 years ago, while the other isolates would be the result of a more recent hybridization. The estimates also suggest that sexual reproduction did not occur every year in all isolates. This is also consistent with

previous findings which indicated that asexual propagation plays an important role in powdery mildew evolution, and that asexual spores might often survive winter on so-called “green bridges” (Wicker et al. 2013).

N. Phylogenetic analysis and divergence time estimates of different *ff. spp.* in *Blumeria graminis*

The divergence times of the different *ff. spp.* are a highly debated topic (Panstruga and Spanu 2014), the various estimations range from 14,000 years to 10 million years ago (Supplementary Table 5) (Wyand and Brown 2003, Takamatsu and Matsuda 2004, Inuma et al. 2007, Oberhaensli et al. 2011, Wicker et al 2013). The higher estimates (4.6 to 11 million years) were all based on an assumed nucleotide substitution rate that is similar to that in plants (Wyand and Brown 2003, Inuma et al. 2007, Oberhaensli et al. 2011, Wicker et al 2013). The broadest and most recent study used synonymous sites in 5,258 genes of *B.g. hordei* and *B.g tritici* and proposed a divergence time of approximately 6.2 million years (Wicker et al. 2013).

In contrast, Wyand and Brown (2003) proposed that *B. graminis* has a much higher nucleotide substitution rate due to its unusual lifestyle. Thus, a much more recent divergence of the different *ff. spp.* only approximately 14,000 years ago could explain incongruence in phylogenies of the pathogens and their hosts: *B.g. avenae* (oat powdery mildew) clusters with *B.g. hordei*, *tritici* and *secalis* in a phylogeny of tub2 and rDNA ITS. Oat belongs to a different tribe of *Poaceae*, the *Avenae*, while rye, wheat and barley belong to the *Triticeae* tribe (Bouchenak-Khelladi et al. 2008). Therefore, Wyand and Brown (2003) proposed that the different *ff.spp.* originated together with agriculture, in the Holocene, around 14,000 years ago.

To test these two hypotheses we used Bayesian phylogenetic methods with 206 single copy orthologous genes from three *B.g. hordei*, five *B.g. secalis*, thirteen *B.g. tritici* isolates and *N. crassa* as a outgroup. We used an independent gamma rate clock model (Lepage et al. 2007) and calibrated the trees setting the divergence time between *B.g. hordei* and *B.g tritici* alternatively to 5.2-7.4 million years ago or 10,000 – 14,000 years ago.

The consensus trees of both analyses show that the isolates of a *f. sp.* cluster together with maximum support. Moreover *B.g. secalis* and *B.g tritici* are sister taxa, while *B.g. hordei* diverged before, this supports the results of Wyand and Brown (2003) and Inuma (2007) (Fig. 3, Supplementary Fig. 38). The divergence time between the *ff. spp. secalis* and *tritici* was estimated

to have occurred between 168,245 and 240,169 years ago (95% credibility interval, calibrated using the estimation of Wicker et al. 2013). In contrast, a divergence time between 638 and 1,280 years (95% credibility interval) resulted when the suggestion of Wyand and Brown (2003) was used as a calibration point. Moreover, when we used the suggestion of Wyand and Brown (2003) to calibrate the trees we obtained a substitution rate higher than one substitution per base per million years. Since we found orthologous genes between *B. graminis* and *N. crassa,* that diverged more than 100 million years ago Prieto and Wedin 2013), we considered the hypothesis of Wyand and Brown (2003) as less likely and instead used the study of Wicker et al. (2013) as a basis for our divergence time estimates.

We repeated the analysis described above including also the six *B.g. dicocci* isolates and found that they cluster as a monophyletic clade with *B.g. tritici* as closest relative. The divergence time between *B.g. dicocci* and *B.g. tritici* was estimated to be between 92,106 and 205,486 years ago (Supplementary Fig. 42). These estimates suggest that *B.g. tritici* is a hybrid between *B.g. dicocci* and another, yet unknown, form of *B. graminis* that diverged from *B.g. dicocci* sometime after the divergence from *B.g. secalis.* However, without sequence data from the second (unknown) parent, it is not possible to determine at what point the hybridization occurred and when the two parents actually diverged.

O. Genes inherited from *B.g. secalis* in all *B.g. triticale* isolates

B.g. triticale differs from *B.g. tritici* in its expanded host range allowing successful infection of triticale in addition to wheat. As described above (Supplementary Note H), *B.g. triticale* contains genome segments that were inherited from *B.g. secalis* which must be at the molecular basis of expanded host specificity to triticale. One possible explanation is that *B.g. triticale* needs specific *B.g. secalis* genes to infect triticale. Therefore, these would have to be present in all *B.g. triticale* isolates. Interestingly, we found only 51 genomic windows with a *B.g. secalis* genotype which were common in all the *B.g. triticale* isolates (corresponding to approximately 0.2 % of the *B.g. triticale* genome). These 51 windows contain only 4 genes (corresponding to 0.01 % of all genes in *B. g. tritici*). Three of these genes are located within the same window, the fourth is on a different sequence contig.

The approach above was very stringent since only polymorphisms were used that are fixed in the *formae speciales*. Thus, we expected that the genomes of the sequenced isolates contain additional

genes that are located in windows that were not assigned to either of the parents with the above described, stringent criteria. Moreover some genes inherited from *B.g. secalis* could have SNPs that we do not observe in the 5 sequenced *B.g. secalis* isolates that are also present in *B.g. tritici*. Such genes would not be detected. To identify additional genes with a *B.g. secalis* genotype in all *B.g. triticale* isolate, we performed a phylogenetic analysis of 4,556 single copy, orthologous genes (Materials and Methods) and identified 64 genes where all *B.g. triticale* isolates clustered with *B.g. secalis* in the respective phylogenetic tree (Supplementary Table 6). We found that 24 of these genes contain mostly SNPs present in our sequenced *B.g. secalis* isolates, but additionally some few SNPs present in *B.g. tritici*, possibly because they might originate from *B.g. secalis* haplotypes not present in our sequenced isolates or originating from gene conversion events. Two of the 64 genes were identified also by the method described above. Therefore in total we identified 66 genes inherited from *B.g. secalis* in all sequenced *B.g. triticale* isolates. Among the 66 genes, there are 6 putative effector genes that might have a role in host specificity determination. However, our transcriptome analysis of *B.g. triticale* isolates on triticale and wheat showed that these genes are not differentially expressed on the two hosts (Supplementary Note P, Supplementary Fig. 45). Functional studies using host-induced gene silencing (Pliego et al 2013) or other approaches will be needed to identify a possible role in determination of host specificity.

As there is no indication that single genes define host specificity based on the analysis described above, it is possible that host range is determined in a quantitative way. The introduction of new plant species resulting from hybridization (e.g. bread wheat and triticale) could have created new potential hosts on which first generation hybrids could survive without the competition of already existing *ff. spp.* We showed that the *B.g. triticale* isolates have between 6.6% and 17.3% of their genome inherited from *B.g. secalis.* Possibly the first hybrid between *B.g. tritici* and *secalis* was less viable on the new host triticale than later variants that back-crossed with *B.g. tritici.* These observations suggest that the capacity to infect triticale is a quantitative effect resulting from interactions between *B.g. secalis* and *B.g. tritici* genomic components in the *B.g. triticale* genome. *B.g. triticale* has the opportunity to further back cross with both *B.g. tritici* and *B.g. secalis*, but natural selection apparently favors *B.g. triticale* strains with genome contribution of about one-eighth of a *B.g. secalis* parent and seven-eighth of a *B.g. tritici* parent.

P. Transcriptomic analysis

B.g. triticale and *B.g. tritici* are the only two *ff. spp*. of *B. graminis* that can grow on more than one host species (triticale, bread wheat and durum wheat for *B.g. triticale* and both wheats for *B.g. tritici*) (Supplementary Note A). This finding opened the opportunity to study differential expression of one *f. sp.* on different hosts. Such an analysis could potentially identify genes that are involved in host specificity. To address this question we designed the following experiment. We analyzed the expression profile of two *B.g. triticale* isolates on two different hosts (the triticale cultivar Timbo and the wheat cultivar Chinese Spring). RNA was extracted two days after infections from the infected leaves. The time point of two days post infection (2 dpi) was chosen because the fungus already forms haustoria on compatible hosts while its growth is blocked before this stage on non-host species (Supplementary Note B). Moreover 2 dpi it is known to be the peak of expression for two effector proteins of *B.g. tritici* (Bourras et al. 2015). Technical details of RNA extraction and data analysis are presented in Material and Methods.

We found only few differentially expressed genes in *B.g. triticale* on the two different hosts. In particular only 25 genes are differentially expressed with an average log2 fold change greater than two (201 if we low the threshold to 1.5 log2) (Supplementary Figs. 43-44). We conclude that there are no major differences in gene regulation after *B.g. triticale* infects two different hosts, but possibly a fine tuning of the expression of few genes. This might indicate that there is no or little induction of host-specific gene expression in *B. graminis*. These findings confirm previous transcriptomic experiments on the *B.g. hordei*. Hacquard and colleagues (2013), sequenced RNA from barley leaves infected with *B.g. hordei* and from immune compromised *Arabidopsis thaliana* mutants. *Arabidopsis* is non-host for *B. graminis* but the inactivation of three components of the immune system (*PEN2*, *PAD4* and *SAG101*) makes *Arabidopsis* susceptible to *B.g. hordei*. Despite the evolutionary distance between barley and *Arabidopsis*, the authors found that the fungal expression program during infection is very robust and stable.

We then analyzed the expression profile of the 66 genes inherited from *B.g. secalis* in all *B.g. triticale* isolates to determine if they showed any specific expression patterns. We found that only one of them was differentially expressed on the two hosts wheat and triticale with average log2 fold change greater than 1.5 (Supplementary Fig. 45). As reported in Supplementary Note O we found that 6 of the 66 genes are putative effectors and, interestingly, two of them are very highly expressed (rpkm > 1000 in all replicates). These genes might have an important role in the infection of *B.g. triticale* independently from the host. Other genes that could be involved in the host range

expansion of *B.g. triticale* are the recombinant genes. These genes are composed of part of the *B.g. tritici* genotype and part of *B.g. secalis* genotype. The analysis of the 29 genes that are recombinant in the two isolates (THUN-12 and T3-8) for which we have transcriptomic data did not reveal any particular pattern. None of these genes was differentially expressed (with log-fold change > 2) on triticale and wheat. *BgtAcSP- 31175* is a gene recombinant in all *B.g. triticale* isolates (although in T3-4 the first part is deleted) and is possibly the result of a recombination between the *B.g. tritici* and the *B.g. secalis* haplotypes (Supplementary Fig. 46). The gene is relatively highly expressed (rpkm from 400 to1600 depending on the replicate and the combination host/ isolate).

We then analyzed differences in expression between two isolates and found more differentially expressed genes here than when we compared expression of the same *B.g. triticale* isolate on two hosts (wheat and triticale). In particular, we found 65 differentially expressed genes with an average log2 fold change greater than 2 and (49 with more than 3 log2 fold change) (Supplementary Fig. 47). We wanted to test whether genes that are inherited from different parents in the two *B.g. triticale* isolates are differentially expressed between the two isolates. We analyzed the expression level of the 1,304 single copy genes for which one of the isolates carries the *B.g. secalis* genotype and the other the *B.g. tritici* genotype. We found that only two of them were differentially expressed between the two isolates with a log2 fold change greater than two (Supplementary Fig. 47). This indicates that the expression level in *B.g. triticale* does not depend on the parent from which the gene was inherited.

Q. Test of evolutionary models with PhyloNet

Analysis of SNP patterns, nucleotide diversity and host ranges suggested that two *ff. spp.* of *B. graminis (B.g. triticale* and *tritici*) originated through hybridization between different lineages. Phylogenetic analysis of orthologous single copy genes also provided evidence for the hybrid origin of *B.g. triticale*. The coalescence theory provides a statistical framework to test these evolutionary hypotheses making use of gene genealogies. For the mathematical background we refer to the literature cited hereafter. We inferred phylogeny of 4,556 single copy genes (Materials and Methods) with sequences from the 5 *B.g. secalis* isolates, the 13 *B.g. tritici* isolates and the 22 *B.g. triticale* isolates. Analysis of tree topologies revealed that in 3,206 trees, *B.g. tritici* and *B.g. secalis* cluster as two separated monophyletic groups. For 1,152 of these genes all *B.g. triticale* isolates cluster with the *B.g. tritici* isolates. If all trees would show the topology of these 1,152 genes, one

would conclude that *B.g. triticale* evolved recently from *B.g. tritici*. This is exactly the conclusion of former studies based on phylogenies of a few genes (Walker et al. 2011, Troch et al. 2012).

In 64 of the 4,556 single copy gene trees we found that all *B.g. triticale* cluster together with *B.g. secalis* isolates (analogously, if all trees had shown that topology, the conclusion would be that *B.g. triticale* evolved from *B.g. secalis*). Finally there are 1,990 trees in which some *B.g. triticale* isolates cluster with *B.g. tritici* and some with *B.g. secalis*. Based on these trees we would propose a hybrid origin or gene flow between the three *ff. spp.* These considerations demonstrate that gene genealogies provide information on the overall species trees.

Degnan and Salter (2005) developed a method to compute the probability of a gene tree based on a given species tree. Yu and colleagues (2011 and 2012) extended it to phylogenetic species networks providing a tool for testing hybridization using gene phylogenies estimated from sequence data. We used PhyloNet (Than et al. 2008) to estimate the Maximum Likelihood (ML) of five different evolutionary hypotheses (Supplementary Fig. 48) which include tree-like evolution of *B. graminis* (no hybridizations), origin of *B.g. triticale* through hybridization, origin of *B.g. tritici* through hybridization, origin of both through hybridization. Because of computational limitation we produced two subsets of isolates and gene trees. We tested two different, non-overlapping sets of isolates (one *B.g. hordei*, two *B.g. secalis*, three *B.g. tritici*, *dicocci* and *triticale*, total of 12 isolates for each dataset) each of them with 10 different sets of 300 randomly selected single copy gene trees (Material and Methods) inferred with RAXML (Stamatakis 20144) using a GTR + GAMMA model (Tavarè 1986, Yang 1993 and 1994). We then used the Akaike information criterion (AIC) (Akaike 1974) the corrected AIC (AICc) and the Bayesian information criterion (BIC) (Schwarz 1978) to select the best model. Details of the different datasets and all results are given in Material and Methods, in Supplementary Tables 7 and 8 and Supplementary Fig. 48. We did not find any differences between AIC, AICc and BIC. For 16 of the 20 datasets (80%) the favored model was the phylogenetic network in which *B.g. triticale* originated through hybridization of *B.g. tritici* and *B.g. secalis*, and *B.g. tritici* itself originated through a hybridization of *B.g. dicocci* and a (yet undiscovered) lineage of *B. graminis*. In the remaining cases the favored model was the phylogenetic network in which *B.g. tritici* and *B.g. triticale* are sister taxa whose ancestor originated through hybridization between *B.g. dicocci* and a yet unknown *f.sp.* Altogether these results show that, based on coalescent theory, the most likely scenario for the origin of *B.g. tritici* is in a hybridization of *B.g. dicocci*. The most likely origin of *B.g. triticale* is as well the result of a hybridization of *B.g. secalis* and *B.g. tritici*.

CHAPTER 3

Reconstructing the Evolutionary History of Powdery Mildew Lineages (*Blumeria graminis*) at Different Evolutionary Time Scales with NGS Data

Fabrizio Menardo[1,] Thomas Wicker[1,2] and Beat Keller[1,2]

[1] *Department of Plant and Microbial Biology, University of Zürich, Zollikerstrasse 107, Zürich 8008, Switzerland*

[2]*Shared last authors*

Submitted to Genome Biology and Evolution

Abstract

Blumeria graminis (Ascomycota) is a fungal pathogen that infects numerous grasses and cereals. Despite its economic impact on agriculture and its scientific importance in plant–pathogen interaction studies, the evolution of different lineages with different host ranges is poorly understood. Moreover, the taxonomy of grass powdery mildew is rather exceptional: there is only one described species (*B. graminis*) subdivided in different *formae speciales* (*ff.spp.*), which are defined by their host range. In this study we applied phylogenomic and population genomic methods to whole genome sequence data of 31 isolates of *B. graminis* belonging to different *ff.spp.* and reconstructed the evolutionary relationships between different lineages. The results of the phylogenomic analysis support a pattern of co-evolution between some of the *ff.spp.* and their host plant. In addition we identified exceptions to this pattern, namely host jump events and the recent radiation of a clade less than 280,000 years ago. Furthermore, we found a high level of gene tree incongruence localized in the youngest clade. To distinguish between incomplete lineage sorting and lateral gene flow we applied a coalescent-based method of demographic inference and found evidence of horizontal gene flow between recently diverged lineages. Overall we found that different processes shaped the diversification of *B.graminis*, co-evolution with the host species, host jump and fast radiation. Our results are an example of how genomic data can resolve complex evolutionary histories of cryptic lineages at different time scales, dealing with incomplete lineage sorting and lateral gene flow.

Introduction

Blumeria graminis (grass powdery mildew) is a fungal pathogen that attacks grass species belonging to the family of Poaceae. It is considered one of the most important fungal pathogens because of its economic impact on cereal crops, especially wheat and barley, and represents a model system to study biotrophic pathogens (Dean et al. 2012). Like other powdery mildews *B. graminis* is an obligate biotroph, depending on living host to complete its life cycle. Grass powdery mildew is considered a single species, the sub-specific taxonomical category *forma specialis* (*f.sp.)* is used to distinguish between forms which show only minimal (or no) morphological differences but are distinct because they occur on different host species (Table 1). Following this definition different *formae speciales* (*ff.spp*) can normally not mate because of the strict host specialization, but they can occasionally mate on alternate hosts (Schulze-Lefert and Panstruga 2011). However,

experimental crosses have been successful only for some of the *ff.spp* (Hiura 1965, 1978; Troch et al. 2014; Menardo et al. 2016).

Table 1. Isolates of *B. graminis* used in this study

Taxon	**Number of isolates**	**Host of origin**[1]
B.g. avenae	1	*Avena sativa*
B.g. dactylidis	1	*Dactylis glomerata*
B.g. hordei	3	*Hordeum vulgare*
B.g. on *Lolium*[2]	1	*Lolium perenne*
B.g. poae	1	*Poa pratensis*
B.g. secalis	5	*Secale cereale*
B.g. tritici1[3]	13	*Triticum aestivum* and *Triticum dicoccoides*
B.g. tritici2[3]	6	*Triticum dicoccoides*

[1]Plants on which the isolates were collected

[2] *B. graminis* growing on *Lolium sp.* was never formally designated as a *f.sp.*

[3] These two groups were found to be genomically divergent despite a common host range (Ben-David et al. 2016; Menardo et al. 2016)

Based on a review of studies that assessed host ranges of different forms of *B. graminis* Troch and colleagues (2014) partially contested the validity of the concept of *ff.spp.*, proposing to retain it only for forms infecting cereals which show a stronger host specialization compared to forms infecting wild grasses. Despite the importance of *B. graminis* as a pathogen and model for research on biotroph life style the evolutionary relationship between *ff.spp.* of *B. graminis* and even the validity of the category of *forma specialis* in evolutionary analysis are topics of intense debate (Panstruga and Spanu 2014). Due to the very simple morphology of *B. graminis* and to the absence of polymorphic traits between different lineages, evolutionary analyses of *ff.spp* are limited to phylogenetic studies performed on molecular data. Most of these studies are based on one or few sequenced genes, and led to discordant results and contrasting interpretations: the results of Inuma and colleagues (2007) based on 4 nuclear genes suggested a pattern of co-evolution between *Blumeria graminis* and its host with some exceptions (i.e. host jumps). However Troch et al. (2014) found in a phylogeny of the β-tubulin that all *ff. spp.* that grow on cereals (including *f.sp avenae*) cluster together despite their hosts being distantly related (oat belongs to the tribe Avenae while rye, wheat and barley belong to the Triticeae). These data suggest a recent origin of *ff.spp*. that infect cereals and contradict the hypothesis of plant-pathogen coevolution. It is known that the analysis of few gene topologies between closely related species or in this case supposedly sub-specific taxa can

lead to contrasting or uninformative results. In particular, gene trees can be different from the species tree because of incomplete lineage sorting (ILS) and lateral gene flow between lineages. A high level of ILS is normally expected between recently diverged lineages while gene flow could be substantial between sub-specific taxa (Rosemberg and Nordborg 2002; Sousa and Hey 2013; Posada 2016).

Recently the genomes of the *ff.spp. hordei* and *tritici* were sequenced and their divergence time was estimated to be approximately 6 Ma based on a molecular clock (Spanu et al. 2010; Wicker et al. 2013), two Myr younger than the estimated divergence between wheat and barley (Middleton et al. 2014) supporting the hypothesis of host tracking, a form of co-evolution in which the speciation of the pathogen is delayed compared to the speciation of the host. However the accuracy of these results depends on a correct assumption of the molecular clock rate. The largest study on *B. graminis* that made use of genome-wide data was conducted by Menardo et al. (2016). In that study we discovered that *B.g. triticale*, a *f.sp.* that can infect the man made crop triticale, originated through a hybridization of isolates of the *f. sp. tritici* and of the *f. sp. secalis.* This finding underlined the importance of reticulate evolution and lateral gene flow in *B. graminis*. Moreover this study identified two distinct lineages of *B. graminis* infecting wheat. In addition, infection tests revealed that one lineage was specific to tetraploid wheat and defined it as the new *f.sp. dicocci*. However, Ben-David and colleagues (2016) found that the isolates of the *f.sp. dicocci* can grow on some hexaploid wheat genotypes that were not tested by Menardo et al. (2016), indicating that the specialization of the two lineages is not complete and that they belong to the same *f.sp.* In this study we will refer to *B.g. tritici1* to identify the lineage named *tritici* in Menardo et al. (2016), to *B.g. tritici2* to identify the lineage named *dicocci* in Menardo et al. (2016) and to *B.g. tritici* when we refer to both lineages.

B. graminis can infect at least 4 different tribes of Pooideae (Bromeae, Triticeae, Avenae and Poaea), however until now all studies based on genome-wide data analyzed mildew growing on a phylogenetically limited set of hosts, barley, rye, wheat and triticale, which are all domesticated plants belonging to the Triticeae tribe. Furthermore all studies based on a large set of mildew isolates from phylogenetically distant host species made use of a small number of molecular markers which resulted in contradictory results, probably due to the poor performance of phylogenetic inference methods in presence of ILS and lateral gene flow between lineages.

Here we used genome sequences of 31 *B. graminis* isolates collected on 8 different plant species belonging to three of the tribes attacked by grass powdery mildew. We aimed to reconstruct the evolutionary relationships between different lineages of *B. graminis* and used phylogenomic methods to infer the species tree, the divergence time between lineages and to identify conflicts between gene trees. Moreover when we found discordance between gene trees we applied a

coalescent based approach to identify the most likely species tree and tested for the presence of lateral gene flow between lineages. We reconstructed the evolutionary history of grass powdery mildew finding evidence for co-evolution between host and pathogen, host jumps and fast radiation. Our results highlight the importance of using a diverse set of methods that can deal with different levels of isolation and divergence between lineages.

Materials and Methods

Sampling

We included in the dataset for the phylogenomic analysis all the currently available powdery mildew genomes (3 *B.g. hordei* isolates: GCA_000151065.1, AOLT00000000 and AOIY01000000; 19 *B.g. tritici* isolates: PRJNA183607 and SRP062198; 5 *B.g. secalis* isolates: SRP062198; one isolate of *Golovinomyces orontii* (*Arabidopsis* powdery mildew): PRJEA50317; one isolate of *Erysiphe pisi* (pea powdery mildew): PRJEA50315; and the reference genome of *Neurospora crassa* as outgroup: PRJNA13841) (Galagan et al. 2003; Spanu et al. 2010; Hacquard et al. 2013; Wicker et al. 2013; Menardo et al. 2016). Additionally we sampled and sequenced powdery mildew on *Poa pratensis* (*f.sp. poae*), on *Avena sativa* (*f.sp. avenae*) on *Lolium perenne* (not formally described as a *f.sp.*) and on *Dactylis glomerata* (*f.sp. dactylidis*). Infected plants of *Poa pratensis* and *Lolium perenne* were obtained from the Agroscope Research station in Reckenholz (Zurich, Switzerland), infected plants of *Dactylis glomerata* were collected in Albisrieden, (Zurich, Switzerland). Single infected plants have been collected and kept in an isolated climate chamber at 20 °C and in 16h light / 8h dark conditions. Spores were collected every 5 days. Oat leaves infected with powdery mildew were kindly provided by Dr. Okon and Prof. Kowalczy (University of Lubklin, Poland). Oat powdery mildew was maintained on detached leaf segments with fresh spores and the infected leaf segments were kept on benzimidazole agar plates at 20 °C, 70% humidity and in 16h light / 8h dark conditions.

DNA Extraction and Sequencing

DNA was extracted from spores as described by Bourras et al. (2015). 125 bp paired-end libraries were created and sequenced with Illumina Hi-Seq at the Functional Genomics Center of Zürich obtaining about 123, 77, 109 and 83 millions of paired reads for *B.g. poae*, *avenae*, *dactiliydis* and *B. graminis* infecting *Lolium*, respectively, all sequences are deposited at the sequence read archive with accession number SRP062198. Bad quality reads were filtered out with sickle 1.33 (Joshi and Fass 2011) with standard parameters. All *de novo* assemblies were performed with CLC Genomic

Workbench 8 with standard parameters and minimum contig size of 200 bp.

Identification of Homologous Genes, Alignments and Gene Tree Discordance Analysis

In a previous study we identified a set of 206 single copy genes that are suitable for phylogenetic analysis in powdery mildew (Menardo et al. 2016). We could retrieve 93 of these genes in all newly sequenced genomes with gmap (version 2013-07-20) (Wu and Nacu 2010), using as template the genes of the *B.g. tritici* reference isolate 96224. The concatenated alignment resulted to be 239,655 bp long and contained a minimum of 91,299 bases for *G. orontii.* All alignments were performed with muscle 3.8.31 (Edgar 2004). In Menardo et al. (2016) we also identified 4,556 homologous single copy genes in the *ff. spp. tritici, secalis, hordeii and triticale*. These genes could not be used in the phylogenomic analysis because they are too divergent or absent in the outgroups *N. crassa, G. orontii* and *E. pisi*. However we retrieved those 4556 single copy genes in the genome of one isolate for each of the lineages of *Blumeria graminis* (96224 for *B.g. tritici1*, 220 for *B.g. tritici2*, the reference isolate DH14 for *B.g. hordei* and S-1201 for *B.g. secalis*, for all other lineages there was only one isolate available). We could find 4,057 of these genes as a single copy in all analyzed genomes and aligned them with muscle 3.8.31 (Edgar 2004). Gene trees were inferred with RAXML 8.0.22 Stamatakis (2014) using a GTR + GAMMA model (Yang 1993, 1994). We used Newick Utilities (Junier and Zdobnov 2010) to identify and summarize topology patterns.

Bayesian Phylogeny and Divergence Time Estimation

With the concatenated alignment (93 genes and 34 taxa), we performed a phylogenetic analysis with MrBayes 3.2.2 (Ronquist et al. 2012) using the independent gamma rate clock model (Lepage et al. 2007). Variation of substitution rates across sites was modeled with a discretized (4 categories) gamma (Γ) distribution (Yang 1993, 1994). The chains have been let free to sample all models of the GTR model family using reversible jump Monte Carlo Markov Chain (Huelsenbeck et al. 2004). We ran 10 independent analyses of 5 million generations each, sampling every 10,000 generations and discarding the first 1,250,000 generation as burn-in. The analysis was repeated several times with different number of generation (5 and 25 million with 25% burn-in) giving very similar results in all the analyses. To calibrate the tree we set 3 calibration points using uniform priors: the first is the divergence between *Letiomycetes* (to whom belong the powdery mildew) and *Sordariomycetes* (to whom belongs *N. crassa*) and was defined as the narrowest range including all the different estimations of Prieto & Wedin (2013) and Beimforte et al. (2014) (160-320 Ma) . For the other calibration points we used the estimated divergence between the hosts of different mildews and set it as the oldest possible age for the divergence of the mildews. This is equivalent to a flat prior for the divergence time spanning from the oldest possible divergence of the host to the present. We set

the divergence between monocot and dicot as oldest possible divergence between *B. graminis* and the dicot powdery mildew *G. orontii* and *E. pisi* (200 Ma, Chaw et al. 2004). The origin of *Pooideae* was used as oldest possible radiation of the *ff. spp*. of *B. graminis* (in particular the split between *B.g. poae* and the other *ff.spp.*) (57 Ma, Bouchenak-Khelladi et al. 2010). All calibration points are secondary calibration points, meaning that they are estimations obtained from trees that were calibrated with fossil age estimations.

Mapping, SNP Calling and Principal-Component Analysis (PCA)

We observed that all isolates of the *ff. spp. tritici, dactylidis* and *secalis* cluster together in the phylogenomic analysis and are much more closely related between them than the other *ff.spp.* We therefore mapped the raw Illumina reads for all isolates of these *ff. spp.* on the *B.g. tritici* reference genome (Wicker et al. 2013) using Bowtie2 (Langmead and Salzberg 2012) with option --score-min L, -0.6, -0.25 (this option allows for approximately four mismatches every 100 bp). We used the following command in SAMtools 0.1.19 (Li et al. 2009) to convert formats and collect information about single genomic positions: view, sort, mpileup -q 15 (only reads with mapping quality greater than 15 were considered). Finally, we used bcftools to generate a VCF file that was parsed with in-house Perl scripts (available upon request). We considered as high-confidence SNPs only positions with a minimum mapping score of 20, a minimum coverage of 8× and a minimum frequency of the alternative call of 0.9. The principal component analysis was performed with the R package GAPIT (Lipka et al. 2012).

Fastsimcoal2

To make a more efficient use of the genome-wide data for the clade that include *B.g. tritici*, *B.g. secalis*, and *B.g. dactylidis* and to test for gene flow between these different lineages we used a method that fits demographic models to the observed multi-dimensional site frequency spectra implemented in fastsimcoal2.5.2.8 (Excoffier et al. 2012). We selected unlinked (at least 2,500 bp apart, average r^2 between adjacent SNPs in *B.g. tritici* = 0.12, computed with VCFtools 0.1.14 Danecek et al. 2011), neutral (at least 2,500 bp far from a gene) SNPs without missing data. After filtering we obtained 19,270 SNPs. The folded site frequency spectrum (FSFS) was calculated with a home-made Perl script (available upon request). Since fastsimcoal2 cannot perform model search we limited our analysis to the three most commonly observed tree topologies. To test for gene flow between the different lineages we modified the three basic models adding one bidirectional introgression between all possible pairs of lineages (6 additional models for each of the basic trees). To distinguish between introgression and migration we modified the basic tree models allowing migration between a pair of lineages for all possible pairs combinations (6 additional models for

each of the basic trees). Finally we tested the hypotheses that *B.g. tritici1* originated through hybridization of *B.g. tritici2* and an unsampled lineage (as proposed in Menardo et al. 2016) implementing this scenario for each of the basic trees (3 additional models). In total, we tested 42 different models. We optimized the likelihood of each model in 100 independent runs, each using 100,000 simulations for every cycle of the conditional maximization algorithm (ECM). We set a minimum of 10 ECM cycles and a maximum of 40, entries in the FSFS with less than 10 observations were not considered in the likelihood computation. Control files of the fastsimcoal2 analysis are provided in Supplementary Material, model comparison was performed with the Akaike information criterium (AIC).

Results

Phylogenomic analysis

To reconstruct the phylogenetic relationships between different lineages of *B. graminis* we used 93 single copy genes (total alignment of 239,655 bp) in a bayesian partitioned analysis. We found that the *ff. spp.* of *B. graminis* cluster as a monophyletic group that diverged from the mildews infecting dicots between 75 and 83 Ma (Fig. 1). The diversification of the *ff. spp.* started between 22.4 and 25.1 Ma with the divergence of *B.g. poae*, supporting the hypothesis of coevolution. Indeed *Poa* was the first lineage to diverge among the grasses infected by lineages of *B. graminis* analyzed in this study, 26-39 Ma (Bouchenak-Khelladi et al. 2008, 2010). Between 12.8 and 14.3 Ma the lineages of *B.g. avenae* and *B. graminis* growing on *Lolium* diverged. In this case the topology of the pathogen is only partially concordant with the topology of the host, the tribe Avenae is the sister taxon of the Poeae tribe and not of the Triticeae to whom *Lolium* belongs. This could be the result of a host jump of *B. graminis* from Triticeae to the Avenae. Between 7.1 and 8.0 Ma *B.g. hordei* diverged from the tritici clade. This also supports the hypothesis of co-evolution, indeed the divergence between barley and wheat was estimated to be about 8 Ma by Middleton et al. (2014). Finally we found a recent radiation between 170,000 and 280,000 years ago that led to the origin of the *ff. spp.* infecting *Secale, Triticum* and *Dactylis* (named tritici clade from now on). *Dactylis* belongs to the Poeae while *Secale* and *Triticum* belong to the Triticeae, suggesting another very recent host jump from Triticeae tribe to Poeae tribe. Overall the phylogenomic analysis supported the hypothesis of co-evolution of some lineages of *B. graminis* and their hosts, but also revealed topological patterns which are compatible with host jumps (*B.g. avenae* and *B.g. dactylidis*) and a recent radiation of the *ff. spp. secalis*, *tritici* and *dactylidis.*

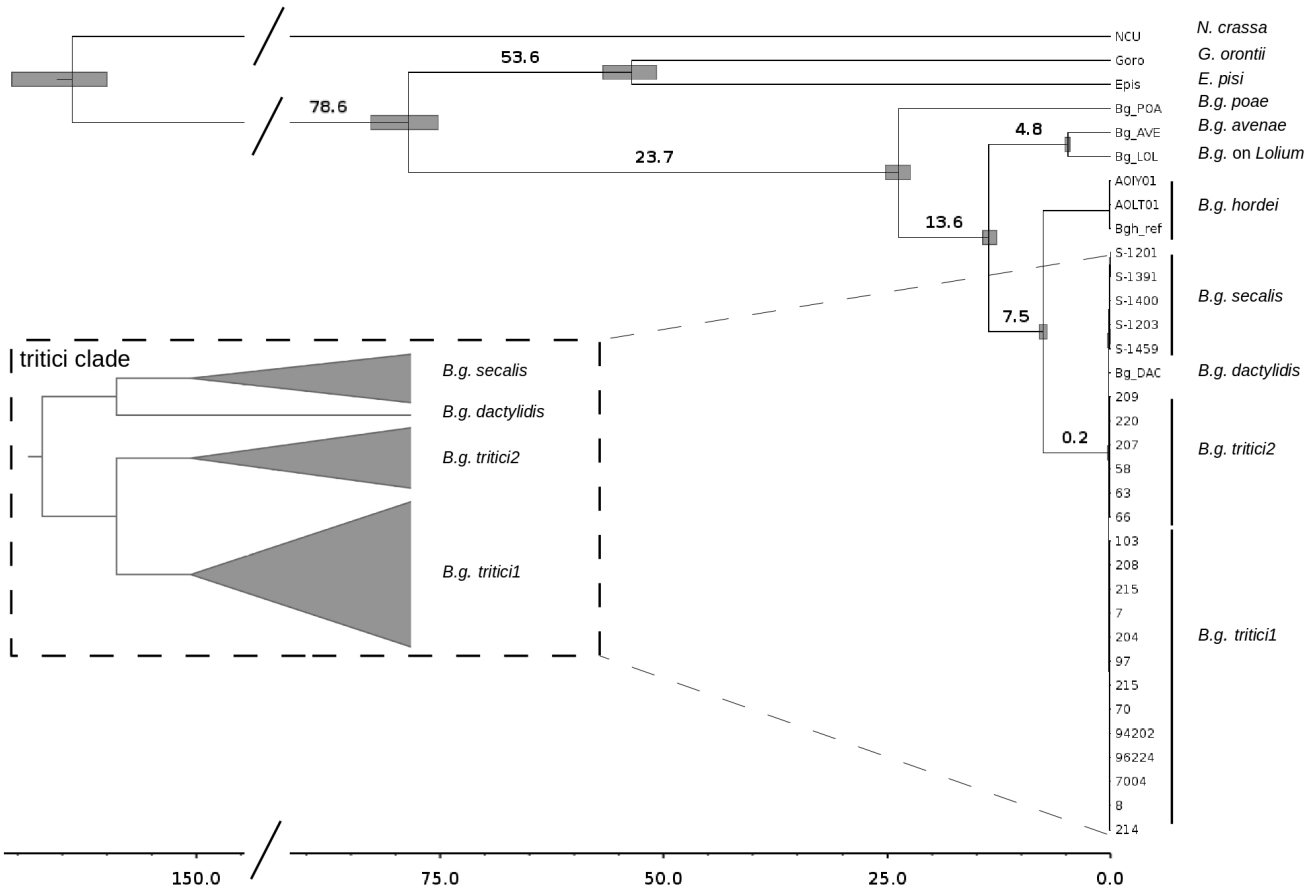


Figure 1. Bayesian consensus tree of powdery mildew strains. Labels on branches represent the median age estimate for the divergence time, gray bars represent the 95% confidence interval of the divergence time, the scale is in million years. In the dashed panel a zoom of the tritici clade is shown (not in scale). All visible ramifications have maximum posterior probabilities.

Genome-wide gene topologies analysis

To investigate the occurrence of incomplete lineage sorting (ILS), reticulate evolution or lateral gene flow in *B. graminis* that could be overlooked by the concatenated phylogenomic analysis we identified 4,057 homologous single copy genes in *B. graminis* and inferred the maximum likelihood tree for each of them singularly using one isolate for each lineage of *B. graminis*. We found 75 different topologies among the 4,057 gene trees with the 14 most frequent topologies observed in more than 95% of the trees. We computed the proportion of gene trees that support the clades observed in the three most frequently inferred topologies (found for 53.7% of genes) and found that the incongruence between topologies are localized in the tritici clade (Fig. 2). Conflicts between gene topologies could be due to ILS, to gene flow or because for some genes there is not a sufficient number of nucleotide changes to robustly reconstruct the gene tree.

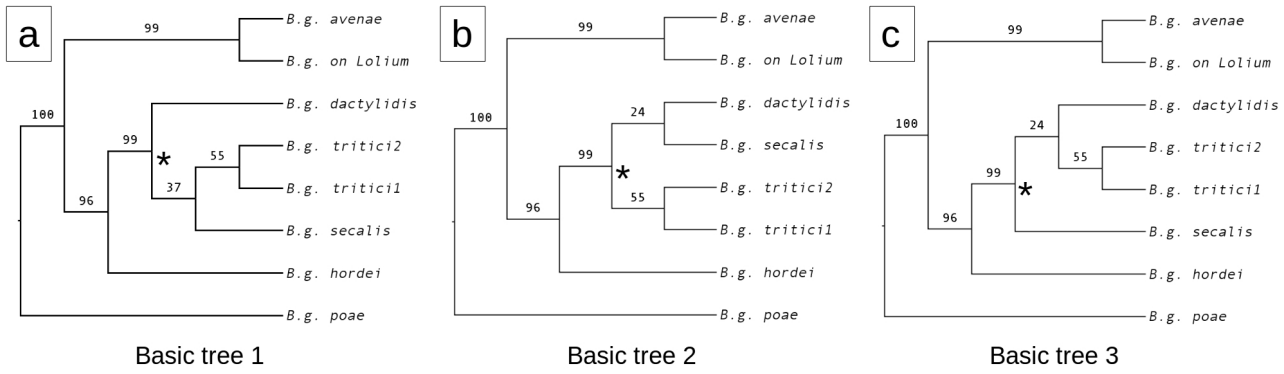


Figure 2. The first (a) second (b) and third (c) most frequently observed gene tree topologies, representing 19.6%, 18.3% and 15.8%, respectively of the 4,057 single copy gene trees. The percentage of single gene tree topologies that support a clade is reported on the relative branch, most of the discordances between tree topologies is localized in the tritici clade (marked with *), in particular regarding the relative positions of *B.g. secalis* and *B.g. dactylidis*.

PCA and demographic inference with fastsimcoal2

Population genomics methods based on SNP data can be more informative on closely related lineages than phylogenomic analysis. We profited from the similarity between the isolates belonging to the tritici clade and mapped the raw Illumina reads on the *B.g. tritici* reference isolate 96224 with comparable overall alignment rate for all isolates (70%). After filtering we obtained 684,338 SNPs, that were used in a PCA and found that isolates of *B.g. secalis*, *tritici1* and *tritici2* cluster in three different groups as described in Menardo et al. (2016). The *B.g. dactylidis* isolate does not cluster with any of these groups indicating that it is part of a different lineage (Fig. 3).
In this study we included all available *B. graminis* genomes except *B.g. triticale* sequences. To test the possibility that *B.g. dactylidis* and *B.g. triticale* are genomically similar and belong to the same lineage we repeated the PCA analysis after addition of two *B.g. triticale* isolates. We found that they cluster in a separate group that does not include *B.g. dactylidis* (Appendix 1).
SNP data can be used to test models for different evolutionary history of closely related lineages.

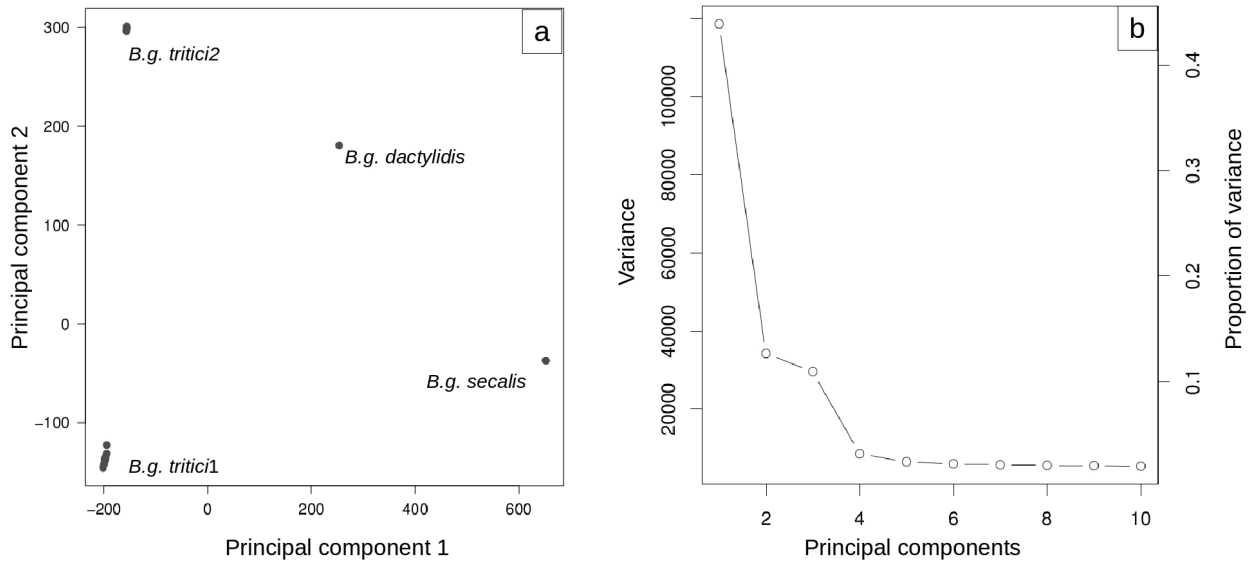


Figure 3. a) Principal component analysis of *B. graminis* isolates based on 684,338 SNPs. There are four distinct groups that correspond to the four lineages of the tritici clade identified with the phylogenomic analysis. **b)** Proportion of the variance explained by different principal components, the first and the second together explain 56.6% of the variance, indicating a high level of genetic structure in the dataset.

We identified 19,270 neutral unlinked SNPs to compute the folded multidimensional site frequency spectrum (FSFS) of the four lineages (*B.g. tritici1*, *tritici2*, *secalis* and *dactylidis*) and used fastsimcoal2 to fit different demographic models to the observed FSFS. Due to the high dimensionality of the models evaluated by fastsimcoal2 and to the computational requirement it is not possible to exhaustively search the model space. Therefore we decided to test the three most frequent gene tree topologies (shown in Fig. 2). We found that the model representing the topology in Fig. 2a (basic tree 1) was the most likely followed by the model representing the topology in Fig. 2c (basic tree 3), and the model representing the topology in Fig. 2b (basic tree 2) being the least likely (Table 2). We then tested for gene flow between lineages. We modified each of the three basic trees adding introgression or migration between all possible pairs of lineages (12 additional models for each basic tree). We found that the majority or all the models with gene flow performed better than basic tree 1, 2 and 3 (9, 12 and 8 models, respectively), and that introgression models generally performed better than migration models (i.e. the best 7 models were introgression models) (Table 2). Additionally we tested the hypothesis that *B.g. tritici1* originated from a hybridization between *B.g. tritici2* and an unsampled lineage based on Menardo et al. (2016).
We implemented this scenario for each of the three basic trees and found that the hybridization models are always more likely than the basic tree models but less likely than other models with

introgression and migration (Table 2). We found that the most likely among the tested model was basic tree 3 with introgression between *B.g. tritici2* and *B.g. secalis*, (basic_tree3_intro12 in Table2 and Fig. 4). Overall these results suggest the occurrence of gene flow between lineages in the tritici clade, especially between *B.g. tritici1*, *B.g. tritici2* and *B.g. secalis*).

Table 2. Results of the fastsimcoal2 analysis

basic_tree1				basic_tree2				basic_tree3			
Model name[1]	logL[2]	N. of parameters[3]	AIC[4]	Model name[1]	logL[2]	N. of parameters[3]	AIC[4]	Model name[1]	logL[2]	N. of parameters[3]	AIC[4]
basic_tree1_intro01	-38198	13	7642	basic_tree2_intro12	-38314	14	7665	basic_tree3_intro12	-37181	13	74387
basic_tree1_intro12	-38227	13	7648	basic_tree2_intro02	-38602	14	7723	basic_tree3_intro02	-38364	13	76753
basic_tree1_intro02	-38367	13	7676	basic_tree2_intro01	-39396	14	7882	basic_tree3_mig12	-39104	12	78233
basic_tree1_mig01	-38909	12	7784	basic_tree2_mig02	-39429	12	7888	basic_tree3_intro01	-39124	13	78275
basic_tree1_mig12	-39120	12	7826	basic_tree2_mig01	-39590	12	7920	basic_tree3_mig02	-39255	12	78533
basic_tree1_H	-39132	13	7829	basic_tree2_mig12	-39635	12	7929	basic_tree3_mig01	-39785	12	79595
basic_tree1_mig13	-39150	12	7832	basic_tree2_intro13	-40017	13	8006	basic_tree3_H	-40006	13	80037
basic_tree1_intro13	-39303	13	7863	basic_tree2_H	-40425	13	8087	basic_tree3_intro23	-40146	13	80319
basic_tree1_intro03	-39316	13	7865	basic_tree2_intro23	-40453	13	8093	basic_tree3_mig23	-40284	12	80591
basic_tree1_mig02	-39342	12	7870	basic_tree2_intro03	-40493	14	8101	basic_tree3	-40306	10	80632
basic_tree1	-39439	10	7889	basic_tree2_mig23	-40658	12	8134	basic_tree3_intro13	-40355	13	80736
basic_tree1_mig03	-39440	12	7890	basic_tree2_mig03	-40786	12	8159	basic_tree3_mig03	-40368	12	80760
basic_tree1_mig23	-39462	12	7894	basic_tree2_mig13	-40935	12	8189	basic_tree3_intro03	-40390	13	80807
basic_tree1_intro23	-39483	13	7899	basic_tree2	-41382	10	8278	basic_tree3_mig13	-40434	12	80891

[1] Models tested with fastsimcoal2 are listed in three columns divided by topology and ranked from the most to the least likely. Basic_tree1 (column 1), basic_tree2 (column 2) and basic_tree3 (column 3) indicate the topologies shown in Figure 2a, b and c respectively. The suffixes intro and mig indicate that introgression or migration was modeled between the two lineages reported at the end of the model name (0 = *B.g. tritici1*, 1 = *B.g. tritici2*, 2 = *B.g. secalis*, 3 = *B.g. dactylidis*), the suffix H indicate a hybridization model in which *B.g. tritici1* originated from a hybridization between *B.g. tritici2* and an unsampled lineage. The best model is highlighted in dark grey; [2] Log-likelihood of the model; [3] Number of parameters of the model; [4] Akaike information criterium.

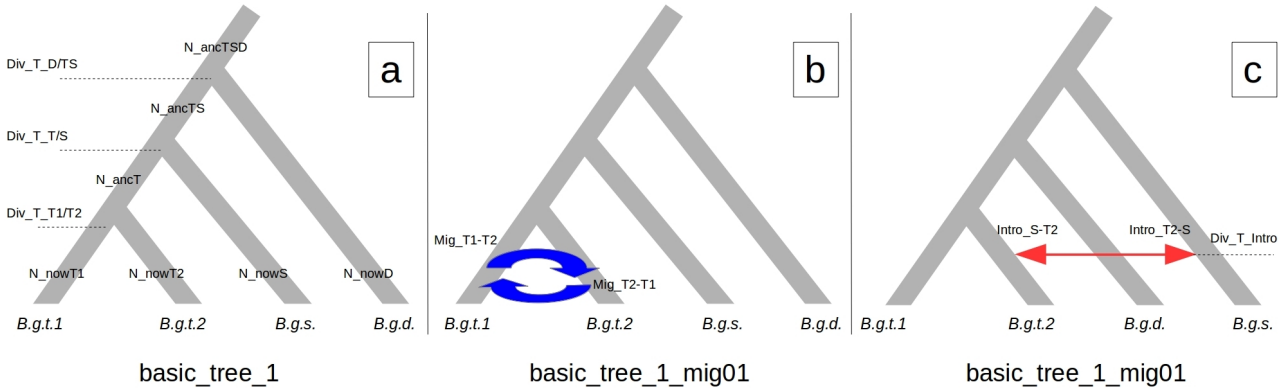


Figure 4. Examples of the demographic models and their parameters used to test different evolutionary hypothesis for the tritici clade. **a)** The model basic_tree1 resulted to be the most likely among the models without lateral gene flow. The parameters of this class of models (models without gene flow, basic_tree1, basic_tree2 and basic_tree3) are also included in the more complex models with migration or introgression. N_now parameters represent the population sizes of the contemporary lineages *B.g. tritici1* (*B.g.t.1*), *B.g. tritici2* (*B.g.t.2*), *B.g. secalis* (*B.g.s.*) and *B.g. dactylidis* (*B.g.d.*). N_anc parameters represent the population sizes of the ancestral lineages. Div_T parameters represent the divergence time between lineages. **b)** The model basic_tree1_mig01 resulted to be the most likely among the models with migration between two lineages. This class of models (models with migration between two lineages) has two additional parameters: Mig parameters represent the migration rate between two lineages. **c)** The model basic_tree3_intro12 resulted to be the most likely among models with introgression between lineages and the most likely among all tested models. This class of models (models with introgression between two lineages) has three additional parameters: Intro parameters represent the proportion of genealogies that move between two lineages at the introgression time, Div_T_intro represent the time of the admixture event.

Discussion

Co-evolution between some lineages of B. graminis and its host

The phylogenomic analysis presented here suggests that some lineages of *B. graminis* co-evolved with their host species. In particular the *ff. spp. poae* and *hordei* and the lineage of *B. graminis* infecting *Lolium* diverged in the same temporal order compared to the divergence of their hosts, *Poa*, *Lolium* and *Hordeum* (Bouchenak-Khelladi et al. 2008) (Fig. 5). The inference of absolute time divergence in the absence of phylogenetically closely related fossils for calibration or accurate estimate of the mutation rate is very challenging (Donoghue and Benton 2007). Unfortunately this is the case for both *B. graminis* and its host. Moreover the estimation of divergence time of *B. graminis* lineages, the divergence of *Poa* and *Lolium* from the Triticeae and the divergence between

barley and wheat are based on three datasets that differ drastically in size (93 nuclear genes, three plastidial genes and whole chloroplast genome respectively) (Bouchenak-Kelladi et al. 2010; Middleton et al. 2014). Therefore, the comparison of divergence time between fungal and plant lineages should be taken with caution. Nevertheless, we found that our estimation of the divergence between *B.g. hordei* and the tritici clade is very similar to the estimation of divergence between their hosts, wheat and barley (7.1-8.0 and 6.0-10.3 Ma respectively) and to the estimation obtained by Wicker et al. (2013) assuming a molecular clock (5.2-7.4 Ma) . The divergence time between *B.g. poae* and the other lineages of *B. graminis* is slightly younger than the divergence time between *Poa* and the Triticeae (22-25 Ma and 26-39 Ma) while the divergence between the lineage of *B. graminis* infecting *Lolium* and the tritici clade is 8 Myr younger than the divergence between *Lolium* and the Triticeae (13-14 Ma and 22-33 Ma). We noticed that the only divergence time estimated by both Middleton et al. (2014) and Bouchenak-Kelladi et al. (2010), between barley and wheat, was overestimated by Bouchenak-Kelladi et al. (2010) compared to Middleton et al. (2014) (8-9 Ma and 10-20 Ma). This could explain the discrepancies in divergence times between lineages of *B. graminis* growing on *Poa* and *Lolium* and their hosts.

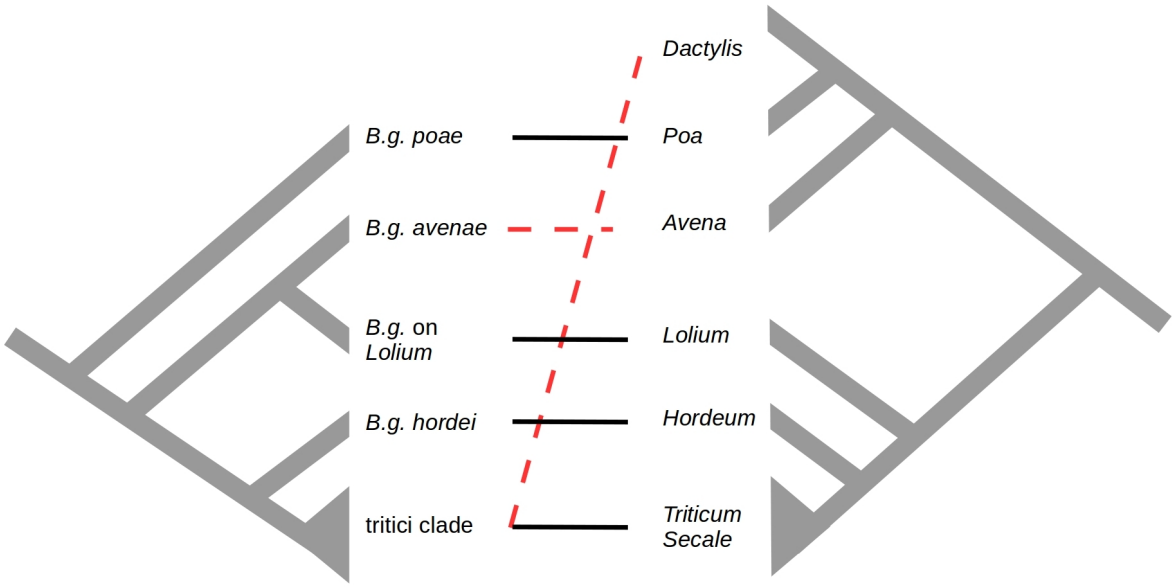


Figure 5. Model of the co-evolution between *B. graminis* and the Pooideae. The phylogenetic tree of *B. graminis* is shown on the left, a simplified version of the phylogenetic tree of the Pooideae is on the right (modified from Bouchenak-Kelladi et al. 2008). Pathogen and host co-evolved (black solid line) with the exception of *B.g dactylidis* in the tritici clade and *B.g. avenae* (red dashed line).

The accuracy of divergence time obtained in this study depends on the correctness of the calibration points, in particular the divergence between Letiomycetes and Sordariomycetes, nevertheless our results are comparable with the estimation of Wicker et al. (2013) based on a molecular clock.

Two major host jumps of B. graminis

Not all lineages of *B. graminis* were found to follow a pattern of co-evolution. Specifically the phylogenetic positions of the *ff.spp. avenae* and *dactylidis* do not correspond to the positions of their hosts (*Avena* and *Dactylis*) on the phylogenetic tree of the Pooideae, suggesting two host jumps (Fig. 5). Moreover the radiation of the tritici clade is very recent (170,000 to 280,000 years ago) and differs from the divergence time of the Triticeae (wheat and rye diverged about 4 Ma). Previous phylogenetic studies, based on four nuclear genes, identified a clade of closely related isolates, which included mildew collected on *Triticum spp.*, *Secale cereale*, *Agropyron spp.*, *Elymus libanoticus*, *Aneurolepidium chinense* and *Aegilops tauschi* (Inuma et al. 2007; Troch et al. 2014). This clade corresponds to the tritici clade observed in this study. Additionally, we found an isolate of the *f.sp. dactylidis* that belong to this clade. This is in contradiction with the study of Inuma et al. (2007) that found that isolates infecting *Dactylis glomerata* clustered outside the tritici clade. Since this was observed for all four genes analyzed by Inuma et al. (2007) it seems possible that there are phylogenetically different lineages that can infect *Dactylis*, and therefore the *f.sp. dactylidis* is probably not a monophyletic group.

Gene flow between recently diverged lineages of the tritici clade

Both the phylogenomic analysis and the PCA revealed a monophyletic group of four lineages that differentiated very recently (less than 280,000 years ago). Using coalescent simulation to fit different demographic models to the observed folded site frequency spectrum we found evidence for lateral gene flow between lineages of the tritici clade. We also found that the tree topology recovered by the phylogenomic analysis is the least likely of all tested models. These findings confirm the importance of considering ILS and lateral gene flow in the reconstruction of evolutionary history of lineages, in particular between closely related ones (Nater et al. 2015). The occurrence of later gene flow is supported by observations on lack of reproductive barriers between *ff. spp.* Several lineages of the tritici clade can mate and produce fertile progeny (*ff. spp. secalis - tritici*, *tritici – agropyri*, Hiura 1965). Moreover Menardo et al. (2016) found that triticale powdery mildew originated from a hybridization of *B.g. secalis* and *B.g. tritici*. On the contrary all attempts to cross lineages of the tritici clade with other older lineages of *B. graminis* failed (Hiura 1978; Troch et al. 2014). We want to point out that the method that we used to infer the probability of different demographic models does not allow for extensive search of the model space. It is possible

that more complex not tested models have a higher likelihood. Moreover, the results of Inuma et al. (2007) suggest that there are additional lineages in the tritici clade attacking different wild grasses which we did not sample in this study. More sequencing efforts, which will have to include multiple isolates for each of the plant host species, are needed to draw a complete picture of the composition and evolution of the tritici clade. It is possible that the analysis of genomics data obtained from a larger sample's set will result in a different interpretation of the evolutionary history of *B. graminis* in general and of the tritici clade in particular. However based on the available data we speculate that the tritici clade is composed of several radiating lineages with different host ranges, these lineages can exchange genetic material between them trough introgression and hybridization. This results in a high potential for the emergence of new pathogens with new virulence spectra.

Formae speciales in *B. graminis* and evolutionary analysis

The classification of *B. graminis* in different *formae speciales* was introduced for the first time by Marchal (1902) and it is used to define "forms" that belong to the same species, are morphologically not distinguishable but infect different plant species (Schulze-Lefert and Panstruga 2011). According to this definition a *f.sp.* does not necessarily represent a distinct evolutionary unit (lineage). However the specialization on different hosts implies, at least in theory, barriers to gene flow between different *ff. spp.* and therefore defines *ff. spp.* as separately evolving lineages which is the only necessary property of a species according to the unified species concept (de Queiroz 2007). However here we focus on the systematic status of *ff. spp.* in *B. graminis* in the light of the genomics data we presented in this paper. We consider first the *ff. spp.* that do not belong to the tritici clade (*hordei*, *poae*, *avenae* and the not formally defined form growing on *Lolium*). These *ff.spp.* represent different lineages that we estimated to be separated by at least 4 million years of independent evolution, even if we cannot exclude the presence of some degree of gene flow between them. This observation together with the evidence for reproductive barriers between these *ff.spp.* (all crosses attempted failed to produce fertile chasmotecia, Troch et al. 2014), suggests that they lack the characteristics of sub specific categories and that it would be more indicated to refer to them as species. Since for these *ff.spp.* we sequenced only one or few individuals and there is evidence that *B. graminis* growing on *Lolium* and on *Avena* are not monophyletic groups (Inuma et al. 2007; Troch et al. 2014) we refrain from modifying the current taxonomy in order to avoid confusion and leave this task to future studies which will use whole genome data from several individuals of each *f .sp.*

Lineages belonging to the tritici clade (*B.g. tritici1*, *B.g. tritici2*, *B.g. secalis* and *B.g. dactylidis*) are more closely related than the others and exchange genetic material between them. Thus, for them a sub-specific classification is justified. However these *ff.spp.* cannot always be considered as

evolutionary lineages because we found two genetically different lineages belonging to the same *f. sp.* (*tritici1* and *tritici2*). We therefore recommend caution when using the concept of *f.sp.* in *B. graminis*. We suggest that this should be strictly limited to define all isolates growing on the same host and no evolutionary implications should refer to this concept.

Conclusions

The advent of next generation sequences provided researchers that study species evolution and attempt to reconstruct the tree of life with a great amount of data. One consequence of this has been the full recognition of the difference between gene trees and species trees and of the processes that cause it (ILS and lateral gene flow) (Posada 2016). These processes have different relevance in different systematic groups and at different timescales in the same group. Our work shows how one can reconstruct evolutionary histories with genomic data using a diverse set of methods that are suited for lineages with a different level of divergence and isolation. The application of these methods to the grass powdery mildew *B. graminis* allowed us to disentangle a complex evolutionary trajectory that includes co-evolution between pathogen and host, host jumps and fast radiations.

Supplementary Material

The alignment and the phylogenetic tree are available as supplementary material. Raw read data have been deposited in the Sequence Read Archive under the accession SRP062198.

Funding

This work was supported by the University Priority Program “Evolution in action” of the University of Zurich and the grant 310030-163260 from the Swiss National Science Foundation.

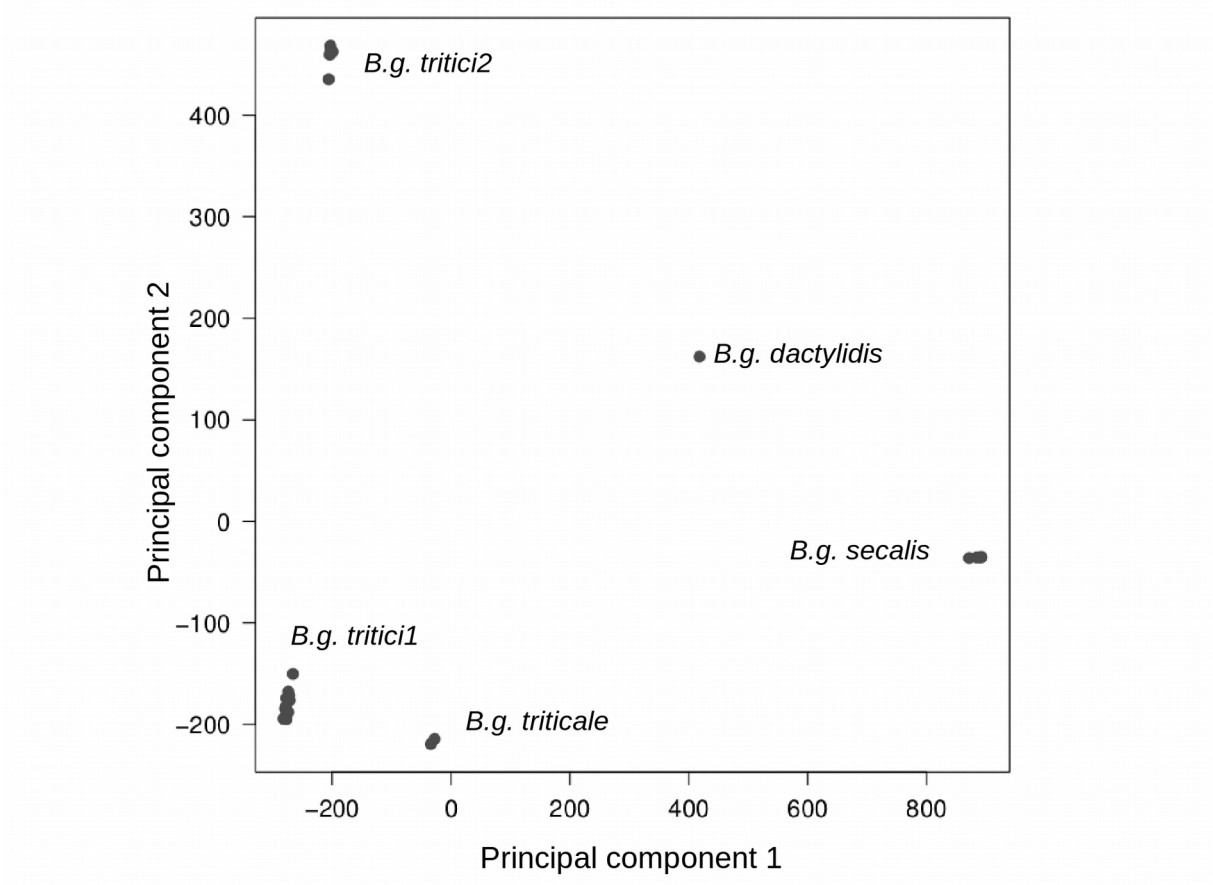
Aknowledgments

We thank Dr. Franz Schubiger (Reckenholz Agroscope research station, Zurich, Switzerland), Dr. Okon and Prof. Kowalczy (University of Lubklin, Poland) for providing plant and fungal material used in this study. We thank the Functional Genomics Center Zurich for performing the sequencing, Linda Lüthi, Helen Zbinden and Geri Herren for technical support. We are grateful to Dr. Bouchenak-Kelladi for providing access on data regarding the phylogeny and divergence time of grasses.

Appendices

Appendix 1

To test the hypothesis that *B.g. triticale* and *B.g. dactylidis* belong to the same genetic cluster we performed a PCA following the same methods as the one used for the analysis displayed in Figure 3, with the addition of two *B.g. triticale* isolates (T1-20 and THUN-12). The results show that the *ff.spp. dactylidis* and *triticale* belong to two distinct groups.



Supplementary Figure 1. Principal component analysis of lineages of the tritici clade and *B.g. triticale*.

CHAPTER 4

Effector analysis in *B. graminis* reveals CSEPs homology to functional domains and fast evolution of ancestral CSEP families driven by gene duplication, gene loss and positive selection.

Part of this work (effector annotation and classification) is included in the manuscript by Praz et al. AvrPm2 encodes a RNAse-like avirulence effector which is conserved in the two different specialized forms of wheat and rye powdery mildew fungus. New Phytologist (2016) in press.

Another part of this work including the evolution of effector families will be part of a manuscript which is in preparation.

Abstract

Most plant pathogens rely on secreted effector proteins to manipulate host metabolism and to evade the plant immune system. Grass powdery mildew (*Blumeria graminis*) is a major cereal disease causing substantial economic losses. *B. graminis* is an obligate biotroph that depends on a living host to conclude its life cycle. The genomes of two *formae speciales* (*ff.sp.*), *B.g hordei* and *B.g. tritici* (mildews of barley and wheat) encode several hundreds of candidate secreted effector proteins (CSEPs). Using a novel approach for *in silico* identification of CSEPs we found 596 CSEPs in *B.g. tritici* and 592 in *B.g. hordei*, 36% and 11% more than previously described. We systematically classified and compared the CSEPs of the two *ff.spp.* and identified known protein domains that were found in the effector repertoires of other pathogenic fungi. To study the evolution of CSEP families we identified CSEPs in five additional forms of *B. graminis* (*B.g. dactylidis*, *B.g. poae*, *B.g. secalis*, *B.g. avenae* and *B. graminis* attacking *Lolium*). We found that most CSEP families are present in all these forms. This indicates an ancient origin of CSEPs in *B. graminis*. Moreover we found that most CSEPs families underwent a fast evolution by a combination of positive selection and multiple gene duplications and losses.

Introduction

Fungal pathogens are the most important cause of disease in cereals and result in considerable yield losses (Deans et al. 2012). Attempts to control fungal disease are generally based either on fungicide treatments or breeding of resistant cultivars. The first defense mechanism that a pathogen encounters during an infection attempt is the recognition of pathogen-associated molecular patterns (PAMPs) by plant membrane receptor, resulting pathogen triggered immunity (PTI). Pathogens can suppress PTI with secreted effector proteins (Jones et Dangl 2006, Giraldo & Valent 2013). Effectors have been suggested to suppress resistance responses, to promote the nutrient uptake and to give structural support to infection structures (Giraldo & Valent 2013, Win et al. 2014). In response plants have evolved another defense strategy that involves cytoplasmic resistance proteins which can recognize effectors and cause an effector-triggered immunity (ETI) which includes defense responses like hypersensitive cell-death (Jones et Dangl 2006). Effector-triggered resistance is frequently overcome by new pathogen strains that according to the classical gene-for-gene model, loose or modify the effectors responsible for the recognition (Jones et Dangl 2006, Stergiopoulos and de Wit 2009). Birch et al. (2006) hypothesized that a large number of effectors with redundant

functions may give an advantage to the pathogen. In particular when one of the effectors is recognized by a resistance protein, the loss or mutation of it would evade recognition, but would not compromise capability to infect the host because of the presence of other genes with the same function. While bacteria usually have between 15 and 30 secreted effectors (Abramovitch et al 2006), several hundred of putative effectors have been identified in the genomes of other pathogens: about 400 in several species of *Phytophthora* (*Oomycetes*) (Tyler et al. 2006, Jiang et al. 2008) and more than 1000 in rust fungi of the genus *Melampsora* (*Basidiomycetes*) (Haquard et al 2012, Saunders et al 2012). In fungi this effector proliferation was often observed in organisms with genomes rich in transposable elements (de Jong et al. 2011).

Grass powdery mildew (*Blumeria graminis, Ascomycota*) is a major cereal pathogen causing yield losses in many parts of the world. It has served as a model to study obligate biotrophic fungal pathogens of plants. The genomes of two *formae speciales* (*ff.spp.*) *B.g. hordei* and *B.g. tritici* (pathogens of barley and wheat) have been sequenced (Spanu et al. 2010, Wicker et al. 2013). In *Blumeria*, candidate secreted effector proteins (CSEPs) have been defined as proteins with a signal-peptide but lacking any trans-membrane domain and homology outside the order *Erysiphales* (Spanu et al. 2010). In total 533 CSEPs have been annotated in *B.g. hordei* and 437 in *B.g. tritici*, corresponding to 7% and 6.7% of the annotated genes (Pedersen et al. 2012, Wicker et al. 2013, Kusch et al. 2014). Most *Blumeria* CSEPs were found to contain a (Y/F/W)XC motif (Godfrey et al. 2010, Pedersen et al. 2012). Evidence of diversifying selection has been shown in *B. graminis* CSEPs (Pedersen at al. 2012, Wicker et al. 2013). Moreover Wicker et al. (2013) compared the genomes of four different isolates of *B.g. tritici* and found that CSEPs are more likely to be lost than other genes. This suggests that in *Blumeria* some effectors could have a redundant function.

It has been proposed that in *Blumeria*, as well as in other fungal pathogens, effectors are secreted through the eukaryotic (Type II) secretory pathway passing through the endoplasmic reticulum (ER) and Golgi vesicle system. However, there is evidence for effectors that might be not secreted or secreted trough a different pathway: two avirulence genes lacking a signal peptide have been described in *B.g. hordei* (Ridout et al. 2006). However, our recent unpublished data suggests that they might not actually be the Avr genes, but near-by transposable element sequences. Additionally, 165 genes unique to *B. graminis*, without signal peptide and with evidence of positive selection have been identified in *B.g. tritici* by Wicker et al. (2013) and were named candidate effector proteins (CEPs). So far functional analysis of *Blumeria* effectors have been performed only with the *B.g. hordei* – barley system. In the absence of a reliable transformation protocol for *Blumeria*, such analysis is challenging. Effector function is usually tested with host induced gene silencing (Nowara et al. 2010, Zhang et al. 2012, Pliego et al. 2013) or by transient expression of the putative effector in single epidermal cells (Schmidt et al. 2014). Several candidate effectors have been confirmed in

the last few years with these two approaches, for some of them it was possible to predict a function based on structural or sequence homology with known proteins (glucanase, peptidase and ribonuclease; Pliego et al. 2013). In other cases the effector target in the plant was identified in yeast-two-hybrid screens (apoplastic resistance proteins PR17 and PR1 in Zhang et al. 2012; thiopurine methyltransferase, ubiquitin-conjugating enzyme and ADP ribosylation factor-GTPase-activating protein in Schmidt et al. 2014; glutatione-S-transferase, malate dehydrogenase and pathogen-related-5 protein in Pennington et al . 2016). Although CSEPs in *Blumeria* were originally identified as secreted proteins without homology outside of *Erysiphales* (Spanu et al. 2010) subsequent annotations included several genes with homology to some functional domain. In particular sequence homology and structural similarity with a ribonuclease domain was found for 79 *B.g. hordei* CSEPs. A major problem of CSEPs identification in *Blumeria* is that the CSEPs annotations were performed in several steps and by different groups in the two *ff. spp. B.g. hordei* and *B.g. tritici* (Spanu et al. 2010, Pedersen et al. 2012, Wicker et al. 2013). This resulted in a lack of homogeneity in the definition and classification of CSEPs in the two *ff. spp.*
In this work we used new criteria for the definition and classification of effectors in *B. graminis* and we used them to determine and compare the CSEPs repertoires of *B.g. tritici* and *B.g. hordei.* Moreover we used the recently sequenced genomes of other lineages of *B. graminis* (*dactylidis*, *secalis*, *poae*, *avenae* and *B. graminis* infecting *Lolium*) to study the evolution of CSEP families in lineages that are diverging for more than 20 million of years. We found that most CSEPs can be classified in families, and that some of these families show homology with protein domains that have been suggested before to be involved in pathogenicity in *B. graminis* and other pathogens. Additionally we found that most CSEP families were already present in the most recent common ancestor of all lineages of *B. graminis* and that the evolution of CSEPs was shaped by positive selection and multiple gene duplication and deletions.

Materials and Methods

B.g. tritici transcriptome assembly

Transcriptomic data from Wicker et al. 2013 were assembled with CLC-workbench6 with standard parameters and resulted in 170,132 contigs. To exclude sequences that are not present in *B.g. tritici* we used all contigs in blast searches against the *B.g. tritici* genome and kept only hits with e-values $< 10^{-6}$. The *B.g. tritici* transcriptome assembly resulted in 16,440 contigs, which were used subsequently.

Annotation of new genes in *B.g. tritici*

We used the *B.g. tritici* gene database (Wicker et al. 2013) as query in blast searches against the *B.g. tritici* genome and kept all hits with at least 50 amino acids in length and 20% identity at the protein level. If previously annotated genes were not found in the aligned region we further analyzed a window of ± 500 bp up- and downstream. In a second step we blasted every window matching the above criteria against the transcriptome and kept hits with more that 95% identity at the nucleotide level. If the alignment had an open reading frame (ORF) we extended the sequence, using the transcriptome as template, until the closest start and stop codons. If in the 5' direction we found a stop codon before finding a start codon we retained the original aligned sequence without extending it in the 5' direction. Results were further parsed to avoid redundant annotations. To exclude repetitive elements we used every new gene as query in blast search against the *Blumeria* and *Triticeae* transposable elements (TE) databases (Bg +PTREP12) (Wicker et al. 2013, http://wheat.pw.usda.gov/ITMI/Repeats/, accessed 01/03/2014), hits with blastp e-value ≤ 10^{-5} were excluded. This annotation produced a large number of gene fragments. We retained only ORFs longer than 50 amino acids and we then performed further manual cleaning excluding all the genes that could be aligned more than 10 times in the genome by gmap (version 2013-07-20, Wu & Nacu 2010). This process was iterated with the newly identified genes. The quality of the resulting annotation depends on the completeness of the transcript assembly. Introns are normally not present in RNAseq data and therefore complete transcript are already spliced, however in case of alternative splicing or in case some intron sequences are present the transcriptome assembly can be fragmented (different exons on different transcripts) and therefore the gene annotation will be of lower quality.

Identification of candidate secreted effector proteins

The protein databases of *B.g. hordei*, *Podospora anserina* (Espagne et al. 2008) and *Neurospora crassa* (Galagan et al. 2003) were downloaded from www.blugen.org, http://podospora.igmors.u-psud.fr/download.php and www.broadinstitute.org/annotation/genome/neurospora/MultiDownloads.html, accessed 01/03/2014). After elimination of proteins with homology in the repeat databases (Bg +PTREP12, blastp e-value ≤ $10^{-5,}$, botinst.uzh.ch/en/research/genetics/thomasWicker/trep-db.html) we retained 9,733 proteins for *N. crassa,* 10,604 *for P. anserina* and 6,011 for *B.g. hordei.* To cluster proteins in families we performed all against all blast searches using protein sequences form *B.g. hordei*, *B.g. tritici*, *N. crassa* and *P. anserina*. Hits with e-value greater than 0.001 and alignment length shorter than 70% of the query length were excluded. Protein families were generated with the markov

cluster algorithm implemented in the mcl software setting the main inflation value to 6 (with option -I 6). This gives a fine-grained cluster and will reduces the average number of proteins in one cluster (Enright et al. 2002). We then defined as secreted genes all genes which encode a predicted signal peptide but not a trans-membrane domain (predicted with signalp 4.1, Petersen et al. 2011). Finally we defined as candidate secreted effector proteins (CSEPs) all proteins belonging to families composed exclusively of *B.g. tritici* and *hordei* proteins (no proteins belonging to *N. crassa* or *P. anserina*) with at least one secreted gene. With this pipeline we identified 1,189 CSEPs (596 in *B.g. tritici* and 592 in *B.g. hordei*). We than reclustered these 1,189 CSEPs with the same pipeline (with option -I 1.4) and found that 1,123 of them belonged to a family of at least 2 proteins. To check for the presence of known functional domains we used all CSEPs in a blast search against the conserved domain database CDD (Marchler-Bauer et al. 2011) http://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi (e-value cut-off =0.01).

Alignment, phylogenetic analysis and test for positive selection

All multiple alignments were performed with Muscle 3.8.31 (Edgar 2004). The nucleotide alignments were retro-translated from the protein alignments using TranslatorX v1.1 (Abascal et al. 2010). We computed phylogenetic trees for every family in the secretome, and included in the alignments the homologous familiy/ies in *B.g. hordei*. Raxml 8.0.22 (Stamatakis 2014) was used to find the maximum likelihood tree using a GTR + GAMMA model, bootstrap support was computed with 100 replications. To test for positive selection in families we estimated the likelihood of the Maximum Likelihood tree under M7 and M8 model (Yang et al. 2000) with Paml 4.8 (Yang et al. 2007) and then used the likelihood ratio test with a null χ_1^2 distribution as indicated in the Paml documentation.

CSEPs annotation and classification in additional *B. graminis* lineages

We used the genome assemblies of *B.g. avenae*, *B.g. dactylidis*, *B.g. poae*, *B.g. secalis* and *B. graminis* infecting *Lolium* (Menardo et al. submitted) to annotate CSEPs in these *ff.spp.* We used the 1,189 CSEPs of *B.g. hordei* and *B.g. tritici* to annotate homologous proteins with MAKER 2.31.8 (Cantarel et al. 2008) and masked repeats present in *B. graminis* repeats database. CSEPs of all *ff.spp.* were then clustered in families with the pipeline described above and phylogenetic trees were inferred with Raxml 8.0.22 (Stamatakis 2014) PROTGTR + GAMMA model, bootstrap support was computed with 100 replications.

Results

Gene annotation

Based on pilot studies for specific CSEPs we suspected that the gene annotation in *B.g. tritici* was incomplete, particularly for CSEPs. To identify any additional genes that were not discovered in previous searches we used the *B.g. tritici* transcriptome data (Wicker et al. 2013) as well as the genome sequence. The *B.g. tritici* genome was used as a database for iterative blast search using all *B.g. tritici* genes as query. Regions with blast hits without annotated genes were used as query in blast search against the transcriptome assembly. Alignments with an open reading frame were extended in both directions, using the transcriptome as template, until the closest start and stop codon. With this pipeline we could annotate expressed genes. In 9 iterations we annotated 731 new genes in *B.g. tritici*. All together the *B.g. tritici* genome is now predicted to have 7,139 genes.

CSEPs identification and classification

To perform comparative analysis between the CSEPs repertoires of *B.g. hordei* and *B.g. tritici* we performed a de novo CSEP prediction in one single analysis for both *ff.spp.* We clustered the protein sequences of *B.g. tritici and B.g. hordei* in families, together with the protein sequences of two non-pathogenic *Ascomycetes: Neurospora crassa* and *Podospora anserina*. *N. crassa* and *P. anserina* have a saprophytic life style and therefore we assume they do not have any genes related to pathogenicity, moreover they are phylogenetically relatively close to *Blumeria* (Prieto and Wedin 2013). We defined as CSEP families all families exclusive to *B. graminis* (no family member in to *N. crassa* or *P. anserina*) and continaing at least 25% of proteins with a predicted signal peptide (signalp 4.1; Petersen et al. 2011). We found 1,189 CSEPs (596 in *B.g. tritici* and 592 in *B.g. hordei*), 1,123 of them (94%) belonged to a family of at least two proteins (Table 1). The new CSEPs estimate increased by 36% and 11% the number of CSEPs in *B.g. tritici* and *hordei* respectively, from 437, as described in Wicker et al. (2013) to 596 and from 533 as described in Pedersen et al. (2012) and Kush et al. (2014) to 592. This increase is due to several reasons:
First, we included the CSEPs proteins without signal peptide but belonging to a CSEP familiy (129 in *B.g. tritici* and 91 in *B.g. hordei*, 81.5% of CSEPs have a signal peptide). This is justified by the observation that often the automatic gene annotation is incomplete or wrong and consequently, the signal peptide can not be predicted by SignalP 4.1 (Petersen et al. 2011). Nevertheless some of these proteins could be truncated forms that are inactive because they lack signal peptide.
Second, we annotated new CSEPs using transcriptomic data in *B.g tritici* resulting in 731 new genes.

Third, we used the protein sequence of *N. crassa* and *P. anserina* to exclude secreted proteins not involved in pathogenicity from CSEPs instead of all sequences outside of *Erysiphales*. In this way we potentially retained CSEPs with homology to some functional domain present in proteins of non-*Erysiphales* organisms.

Table 1. Overview of CSEP families in *B.g. hordei* and *tritici*

Family	Proteins[1]	SP[2]	Bgt[3]	Bgh[4]	Length[5]	Conserved domains (pfam family name)		
1	168	67	88	80	291	1 Aldose_epim	1 zf-DNA_Pol	1 FNR_like
2	78	74	33	45	109			
3	77	59	21	56	299	20 microbial_RNases	1 Na_K-ATPase	
4	76	67	34	42	144	1 vWFA	1 RE_HpaII	1 Esterase_lipase
5	46	43	30	16	109	1 SATase_N		
6	43	34	19	24	309	11 microbial_RNases	1 Dcm	
7	30	15	20	10	264	4 microbial_RNases		
8	27	21	15	12	145	1 microbial_RNases		
9	26	7	17	9	264	2 microbial_RNases		
10	26	26	14	12	116			
11	24	20	6	18	137			
12	20	19	13	7	161			
13	16	16	8	8	123	3 microbial_RNases	1 OCIA	
14	16	14	6	10	299	1 ARS2	1 PHA03160	1 Bindin
15	15	14	4	11	147			
16	14	14	5	9	109			
17	13	9	9	4	310			
18	13	8	6	7	241	12 DUF3129		
19	13	13	5	8	161	4 ML	1 Urease_gamma	1 SPRY
20	13	5	9	4	278			
21	12	11	4	8	122			
22	11	11	7	4	115			
23	11	11	6	5	95			
24	10	10	4	6	127	1 DUF4412		

[1]Number of proteins belonging to the family
[2]Number of proteins with predicted signal peptide
[3]Number of family members in *B.g. tritici*
[4]Number of family members in *B.g. hordei*
[5]Average length of family members in amino acids

We used the conserved domain database CDD (Marchler-Bauer et al. 2011) to identify known functional domains in the CSEPs repertoire. We found that 95 *B.g. hordei* and 70 *B.g. tritici* CSEPs have homology with at least one protein domain. The most represented domain was the RNAse domain (48 genes in 13 families) followed by DUF3129 (domain of unknown function) (11 genes in one family), the esterase-lipase domain (8 genes in 4 families), the MD2-related lipid-recognition domain (ML) (6 genes in 2 families), the histidin-phosphatase domain (HP) (6 genes in 2 families) and the fungal cysteine rich domain (CFEM) (6 genes in 3 families). The RNAse domain was earlier descibed in *B.g. hordei* CSEPs (Spanu et al. 2010, Pedersen et al. 2012) while the other domains have not yet been identified in *B. graminis* CSEPs. However, the ML domain, DUF3129 and the CFEM domain were found to be enriched in putative effectors of rust fungi (Saunders et al.

2012). The combination of these results with the RNA expression analysis performed by Praz et al. (in preparation) on three *B.g. tritici* isolates defines two major classes of effector families: one composed of large proteins (more than 200 aa) with low expression level and one of shorter proteins with higher expression (Fig. 1). The same two classes of CSEPs were already identified by Pedersen et al. (2012) in *B.g. hordei* using protein size, strength of positive selection and differential expression between epiphitic and haustorial structures. Families with homology to RNAse belong mostly to the class of larger proteins (families 3,6,7 and 9) while known Avrs and Svrs (Bourras et al. 2015, Lu et al. 2016, Praz et al. 2016) to the highly expressed and shorter families (families 5, 10 and 12). It is likely that the differences in size and expression level of these two class of CSEPs reflect different biological functions.

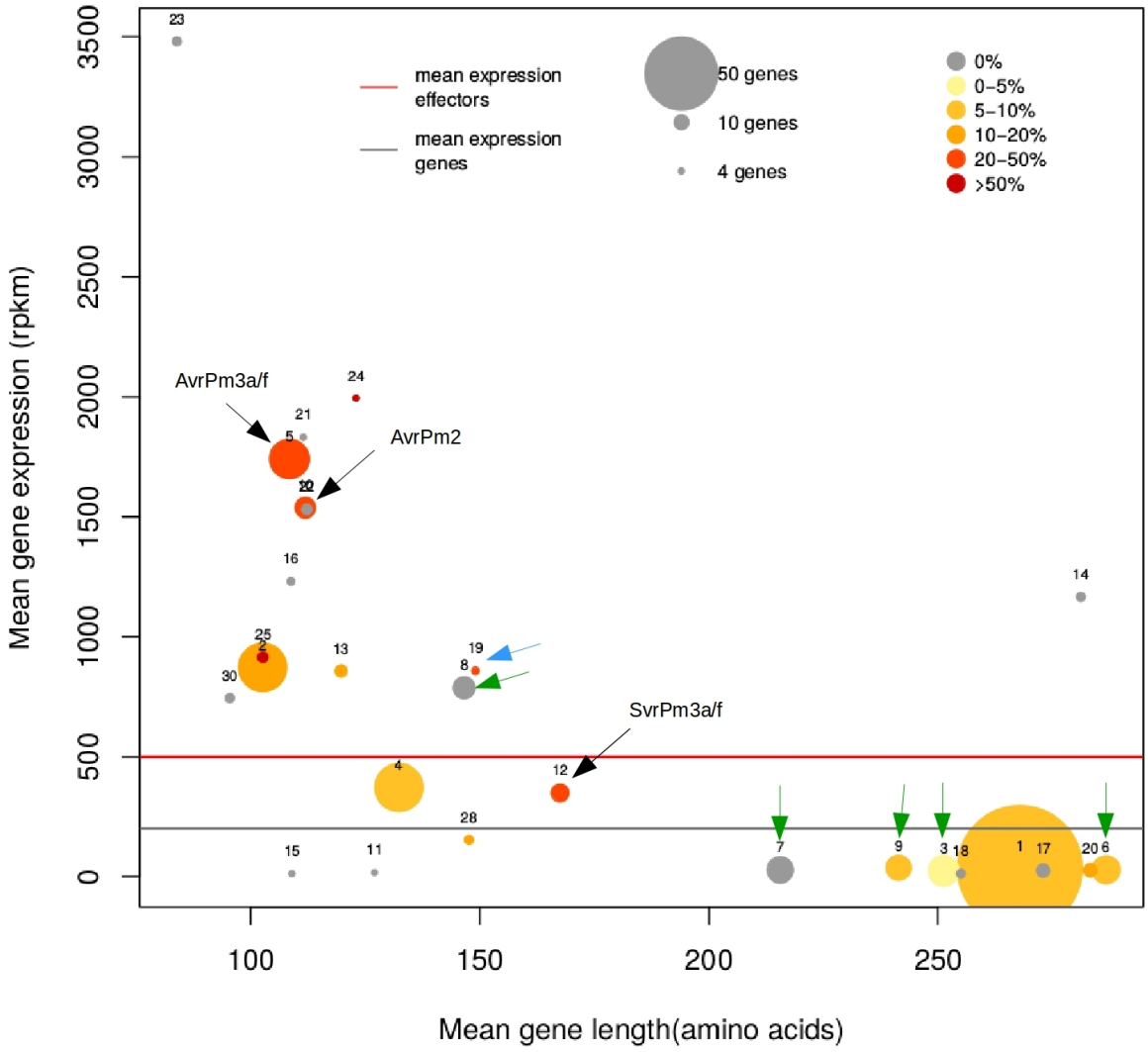


Figure 1. Plot of average protein length and average expression (Praz et al. 2016) for the 24 largest *B.g. tritici* CSEP families. The size of the dots is proportional to family size, the color of the dots indicates the percentage of differentially expressed genes among the three *B.g. triticici* isolates analyzed by Praz et al. (in preparation). Green arrows indicate families with homology to RNAse, the blue arrow indicate the family with homology to the ML domain. Black arrows indicate the families which include known Avrs and Svrs.

Families of large proteins mostly have low expression level on the right of the graph, while on the left there are families composed of short proteins with higher expression level.

More than 95% of CSEPs harbor a Hy(X)(X)C(S) motif

We visually inspected alignments of *B. graminis* families for the conserved (Y/W/F)XC motif identified by Godfrey et al. (2010). We analyzed the 24 families with 10 or more proteins (798 proteins in *B.g. tritici* and *hordei*) in *B. graminis* and despite the general lack of sequence conservation we identified a HyXC motif (Hy is a hydrophobic amino acid, X is a not conserved amino acid) or variants of it (HyCS or HyXXC) in 22 gene families comprising 97% of analyzed CSEPs (Fig. 2). The motif was always present in the first 50 amino acids after the signal peptide. All the different motifs can be generalized as Hy(X)(X)C. The previously described (Y/W/F)XC motif (Godfrey et al. 2010) is a subtype of the Hy(X)(X)C motif.

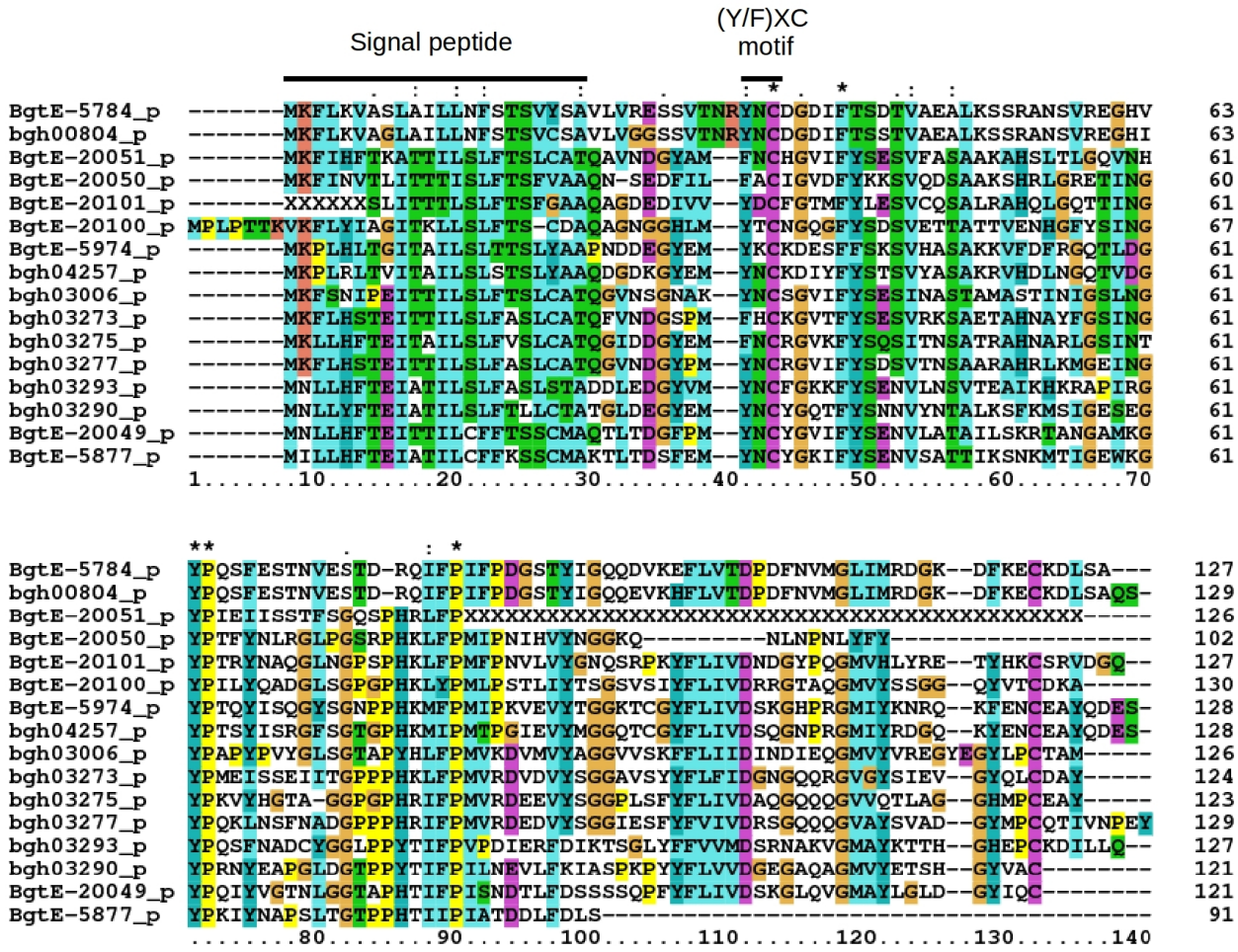


Figure 2. Protein alignment of the CSEPs family 13, amino acids are colored with the clustal code (Thompson et al. 1994). This family shows typical characteristics of all CSEPs family alignments, in particular the low sequence conservation between members. The (Y/F)XC motif and the predicted signal peptide are highlighted.

Analysis of CSEPs evolution

To study the evolution of CSEP families we extended our analysis to other lineages of *B. graminis.* In particular anciently diverged lineages could give information on the presence of CSEPs in the most recent common ancestor of *B. graminis* lineages and help to infer the evolutionary history of CSEP families (chapter 3). Therefore we used the Maker software (Cantarel et al. 2008) to annotate CSEPs in 5 additional forms of *B. graminis* (the *ff. spp. secalis*, *dactylidis*, *poae* and *avenae* and *B. graminis* growing on *Lolium*) using *B.g. tritici* and *hordei* CSEPs as a template. We found a minimum of 361 CSEPs in *B. graminis* growing on *Lolium* and a maximum of 910 CSEPs in *B.g. dactylids.* This large variation in annotated CSEPs could be due to the fact that *B.g. dactylidis* and *secalis* diverged very recently from *B.g. tritici* and *hordei* compared to the other lineages and therefore they have a greater sequence similarity to *B.g. tritici. B.g. secalis* and *dactylidis* have more CSEPs compared to *hordei* and *tritici*, probably because gene annotation in *B.g. tritici* and *hordei* was based on transcriptomic data collected at 4, 8, 12, 24 and 48 hours after infection, therefore genes not expressed during these life stages of the isolate for which the transcriptome was sequenced are not included. The annotation performed for the other lineages is only based on genomic data, therefore not expressed genes are included.

Table 2. Number of predicted CSEPs in different lineages of *B. graminis*

Lineage	Number of predicted CSEPs
B.g. hordei	592
B.g. tritici	596
B.g. secalis	765
B.g. dactylidis	910
B.g. poae	480
B.g. avenae	370
B. graminis growing on Lolium	361

The classification in families showed that 18 of the 24 largest families defined in *B.g. tritici* and *hordei* were found in all analyzed lineages. This finding implies that most CSEP families are present with multiple members in all *B. graminis* lineages and therefore were already present in the most recent common ancestor of *B. graminis* lineages. This suggests that the role of most CSEPs is conserved through all lineages of *B. graminis* which infect different plants, and is not specific for a single host.

In chapter 3 we studied and reported the evolutionary history of lineages of *B. graminis* and found that the first lineage to diverge was *B.g. poa* (about 23 Ma), followed about 14 Ma by a lineage that later originated *B.g. avenae* and *B. graminis* infecting *Lolium* (about 5 Ma). Finally *B.g. hordei* and the tritici clade diverged about 8 Ma. The tritici clade includes *B.g. tritici*, *B.g. secalis* and *B.g. dactylidis* which diverged less than 300,000 years ago. We also showed that single copy gene trees mostly corresponded to the species tree (except for relationship between lineages of the tritici clade) (chapter 3). We checked if this is the case also for CSEPs belonging to the 10 largest families (597 CSEPs). We found only 7 CSEPs that have homologous proteins in all *B. graminis* lineages and for which the gene tree corresponds to the species tree (Fig. 3). In most cases gene duplications and gene deletions altered the gene trees and it was impossible to identify a single orthologous CSEP in all lineages (Fig. 4). Some CSEPs experienced lineage specific multiple gene duplications resulting in clades of several CSEPs which are present in only one lineage of *B. graminis* (up to 69 *B.g. poae* CSEPs in family 5; Fig. 5). We then tested for positive selection acting of CSEP families. Since this analysis can be negatively influenced by false annotations we included only *B.g. tritici* and *B.g. hordei* genes for which there is a more accurate annotation. We found that all of the 24 largest families are under positive selection. Likelihood ratio tests (between M7 and M8 models) for positive selection on the complete families were positive for all tested families (p-value < 0.01), confirming the results of Pedersen et al. (2012) and Wicker et al. (2013). These findings show that most of the CSEPs families experienced a fast evolution by a combination of two different processes, the first is positive selection, acting at the sequence level and probably driven by selection to escape recognition by plant resistance proteins and/or to keep up with modifications of effector targets. The second is based on expansion and contraction of CSEP families. Indeed we found that CSEP families can increase or decrease considerably in gene number in specific lineages in a relatively short evolutionary time.

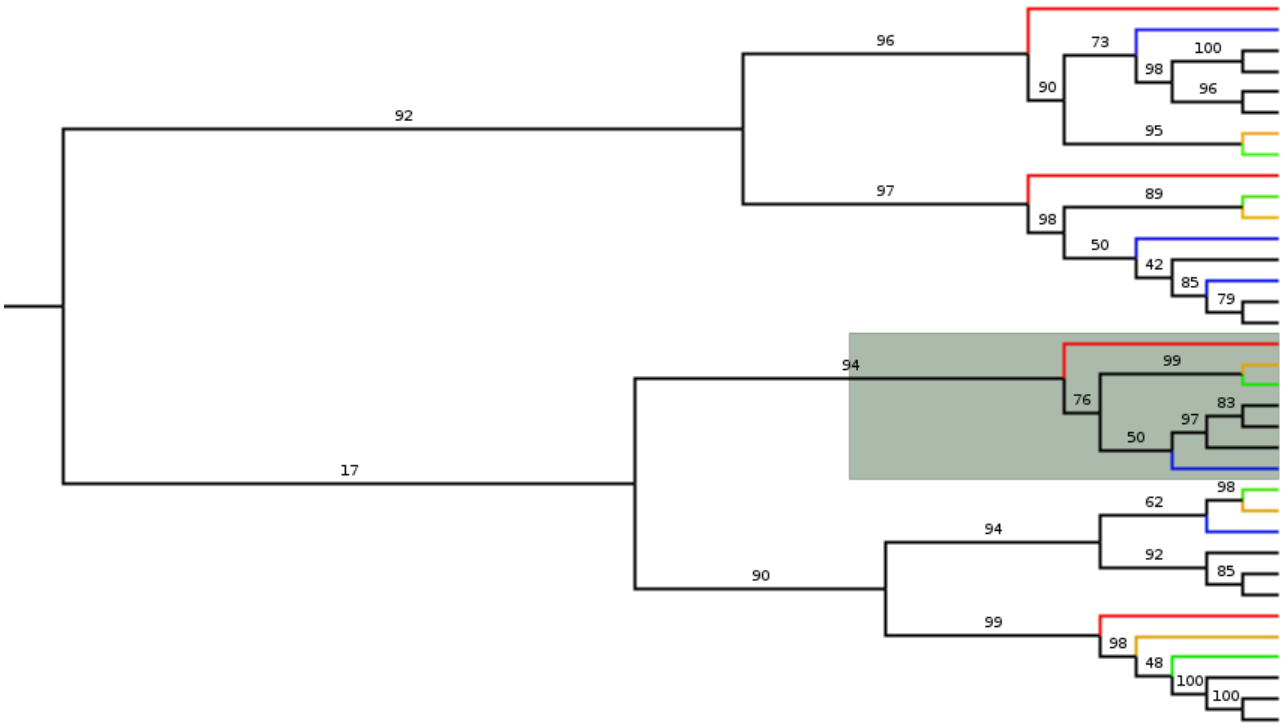


Figure 3. Subclade of the ML phylogenetic tree of CSEPs family 1 (branch lengths are not in scale). A color code identifies the different lineages (red = *B.g. poae*, green = *B. graminis* infecting *Lolium*, orange = B.*g. avenae*, blue = *B.g. hordei*, black = tritici clade (*B.g. tritici secalis* and *dactylidis*), numbers on branches represent bootstrap support. This example shows the evolution of 5 CSEPs from the most recent common ancestor of *B. graminis* lineages to the actual lineages. However only one of them (highlighted in grey) has exactly one homologue in each of the *B. graminis* lineages and a gene tree topology that corresponds to the species tree. All the others have duplications or deletions in some lineages and do not correspond to the species tree.

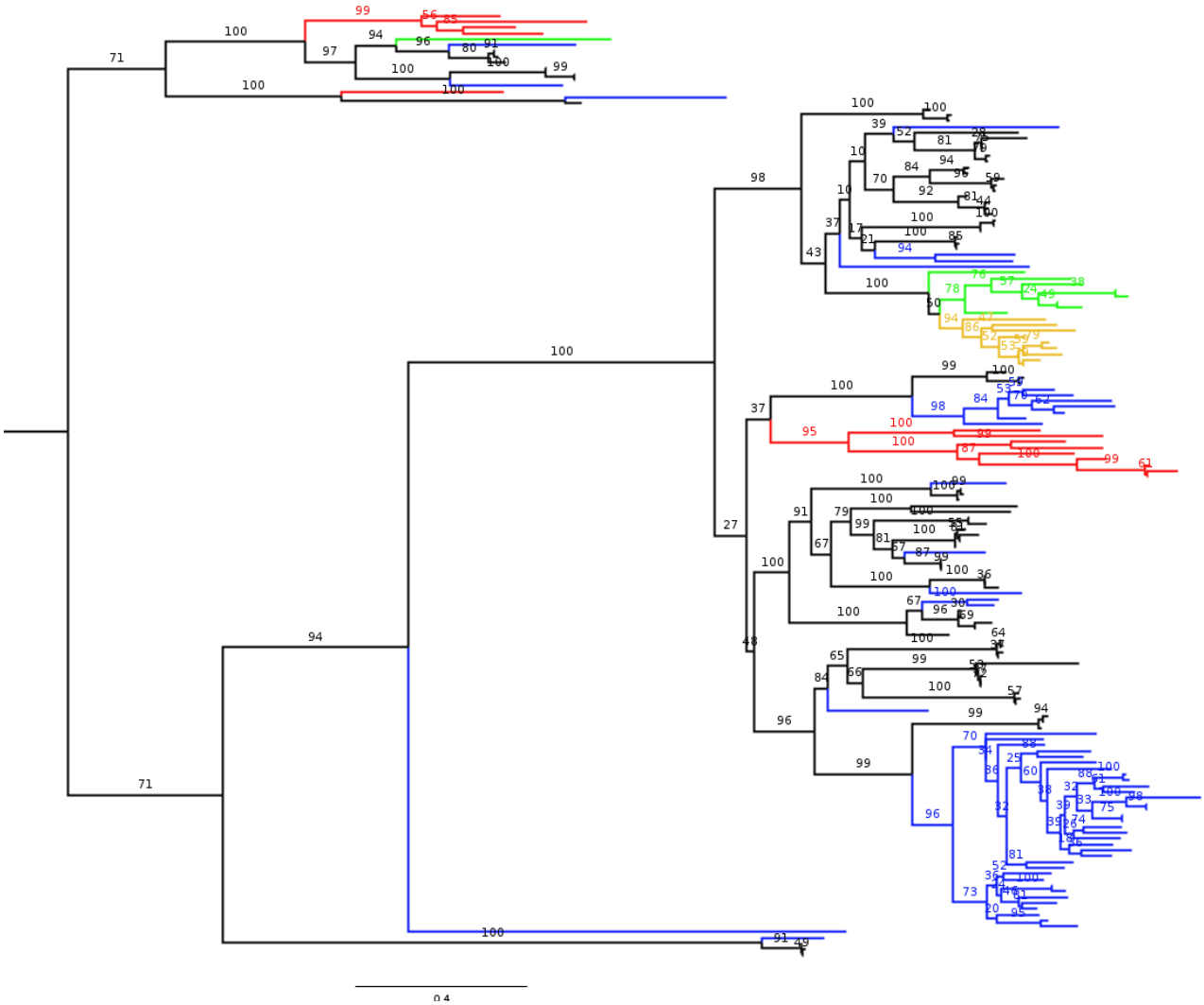


Figure 4. ML phylogenetic tree of CSEPs family 4. A color code identifies the different lineages (red = *B.g. poae*, green = *B. graminis* infecting *Lolium*, orange = B.*g. avenae*, blue = *B.g. hordei*, black = tritici clade (*B.g. tritici secalis* and *dactylidis*), numbers on branches represent bootstrap support. The evolution of this family was heavily shaped by gene duplications and deletions. Particularly evident are lineage-specific expansions where a subclade of the family expanded in only one of the lineages. The scale is in expected substitution per site.

Figure 5. ML phylogenetic tree of CSEPs family 5. A color code identifies the different lineages (red = *B.g. poae*, green = *B. graminis* infecting *Lolium*, orange = *B.g. avenae*, blue = *B.g. hordei*, black = tritici clade (*B.g. tritici secalis* and *dactylidis*). This is the most extreme example of lineage-specific expansion (69 CSEPs in *B.g. poae*). The scale is in expected substitution per site.

Discussion

A short, highly conserved motif in most CSEPs

Among the families with 10 or more genes in *B. graminis* (representing almost 70% of the total CSEPs) 22 of 24 (representing 97% of the analyzed CSEPs) have a conserved motif Hy(X)(X)C(S). A particular variation of the Hy(X)(X)C(S) motif, (Y/F/W)XC was found in effectors of fungal pathogens belonging to *Puccinia* and *Melampsora* (Godfrey et a. 2010, Saunders et al. 2012) and is also the most represented in *B. graminis hordei* and *tritici*. The prominence of the Hy(X)(X)C(S) motif in *Blumeria* CSEPs suggests a possible role of this sequence in delivering CSEPs into the host cell. Another short conserved motif (RXLR-dEER) present in hundreds of effectors of several *Oomycete* species has been shown to be sufficient and necessary for the translocation of such effectors into the host cytoplasm (Dou et al. 2008). Kale et al. (2010) showed that RLXR effectors bind phosphatidylinositol-3-phosphate (PI3P) and enter into the cytoplasm via lipid-raft mediated endocytosis. However the experiments conducted by Yaeno et al. (2011) contradict these results and a clarification of the situation is needed (Ellis and Dodds 2011).

Conserved protein domains in CSEP families

We found that several CSEP families have homology with conserved protein domains. The most abundant was the RNAse domain that was already described in Pedersen et al. (2012). Pedersen and coworkers also showed that using template based modeling of CSEPs it is possible to identify a structural homology to the RNAse domain for several additional CSEP families in *B.g. hordei*. However they also showed that the catalytic residues responsible for the RNAse activity are not conserved in *B.g. hordei* CSEPs. We found three additional protein domains highly represented in the *B. graminis* CSEPs: DUF3129, ML and CFEM domains. These domains have also been found to be enriched in the secretomes of two rust fungi (*Melampsora larici-populina* and *Puccinia graminis f.sp. tritici*) (Saunders et al. 2012). These two rust pathogens are *Basidiomycetes* with an obligate biotrophic life style and they both form intracellular haustoria. The finding that these domains are present in the effector sets of the distantly related pathogen species rust and mildew suggests a convergent evolution of different organisms with a similar life style. DUF3129 (domain of unknown function) and CFEM (cystein rich fungal domain) have not been functionally characterized. The ML (MD2-related lipid-recognition) domain is present in many animal, plant and fungal proteins and is well studied. Proteins with this domain are usually shorter than 200 amino acids and involved in regulating lipid metabolism. In addition, they act as cofactors in recognition of pathogen associated lipids and in the phospholipid transfer through membranes (Inohara & Nuñez 2002). The ML domain could be involved in the interactions occurring between the pathogen

and host membranes. The well characterized protein from which the domain takes its name is MD-2. This protein interacts with Toll-like-receptor 4 and it is necessary to recognize bacterial lipopeptides and lipopolysaccharides (LPS) in animals (Miyake 2007). Structural homology between some plant receptor-like kinase (RLK) and animal Toll receptors suggest similarities in the modes of action (van der Biezen and Jones, 1998) and the receptor-like kinase Nt-Sd-RLK was shown to respond to bacterial LPS in *Nicotiana tabacum* (Sanabria et al. 2012). Among fungal membrane components, so far only ergosterol has been observed to be specifically recognized by plants (Granado et al 1995) causing a typical elicitation response not caused by plant or animal sterols (Rossard et al. 2010). Ergosterol is a perfect PAMP candidate because is not present in plant membranes and can be recognized as non-self by the plant immune system (Klemptner et al. 2014). It is not known how plants recognize ergosterol, but one Arabidopsis RLK was shown to be responsive to others steroids, brassinosteroids and directly bind brassinolide (Wang et al. 2000). Additional protein domains found in *B. graminis* CSEPs are the esterase-lipase domain and the HP (histidine-phosphatase) domain. These two are enzymatic domains involved in a multitude of different processes and it is difficult to precisely relate them with pathogenicity.

Patterns of CSEPs evolution

We showed that most CSEP families are present in all lineages of *B. graminis*. Moreover it was shown that most *B. graminis* CSEPs do not have homologs in the genomes of the *Erysiphe pisi* (pea powdery mildew) and *Golovinomyces oronti* (*Arabidopsis* powdery mildew) (Wessling et al. 2014). These findings indicate that most CSEPs of *B. graminis* originated in the period between the divergence of *B. graminis* from other mildews and the beginning of the differentiation of the different *B. graminis* lineages. In chapter 3 we estimated this time period to be between 80 and 20 Ma. Additionally we found that CSEP families are highly dynamic and in most cases it is impossible to identify direct orthologs across lineages of *B. graminis*. This is due to the gene duplications and losses which acted massively on most CSEP families in all lineages of *B. graminis*. CSEP clades that underwent lineage-specific expansion could be important pathogenic factors specific for one lineage and its host. Similar to our study, Pendleton et al. (2014) found that, in the rust fungus *Cronartium quercuum f. sp. fusiforme* (*Basidiomycetes*), putative effectors families underwent species-specific gene duplications. Moreover Jiang et al. (2008) and Goss et al. (2014) reported repeated effector duplications in *Phytophtora ramorum* (*Oomycetes*), a pathogen with a broad host range. Unequal crossing over was proposed to be the mechanism for repeated gene duplications that led to CSEP family expansions in powdery mildew. This was based on the observation that genes belonging to the same family occur in clusters in the genome (Pedersen et al. 2012). Positive selection also drives CSEP evolution, and we found that all the *Blumeria* CSEP

families are under positive selection, confirming the results of Pedersen et al. (2012) and Wicker et al. (2013). Evidence of positive selection has been found in effectors of different eucaryotes (Wie et al. 2007, Goss. et al. 2014, *Oomycetes;* Hacquard et al. 2012, *Basidiomycetes;* reviewed in Ma and Guttman 2008), and it is driven by the pressure to evade recognition and/or to keep up with modification of effector targets. We found that positive selection acts in parallel to gene duplications. Thus, positive selection and multiple gene duplications and losses are responsible for the fast evolution that we observe in CSEPs of *Blumeria* as well as of other pathogens. The first contributes to the fixation of non-synonymous mutations, the second has two effects: it creates the effector repertoire on which positive selection can act (gene duplication), but at the same time destroys genes, contributing to the increase the diversity of CSEP repertoires in pathogen populations.

CHAPTER 5

Outlook

Neutral evolution of *B. graminis*

The availability of genomic sequences for most of the *ff.spp.* of *B. graminis* allowed a detailed reconstruction of the evolutionary history of these lineages. Often the results of this work were in contrast with previous studies which were not based on genome wide analysis. This shows how inferring species evolution from analysis based on few molecular markers requires a lot of carefulness and awareness of the assumptions that one is making (often these assumption are absence of gene flow and incomplete lineage sorting).

The most important finding of this PhD work was that the recently emerged powdery mildew growing on triticale is a hybrid between the two forms growing on wheat and rye (chapter 2), while previous studies proposed that it evolved from the wheat powdery mildew, based on the phylogeny of a few genes (Walker et al. 2011, Troch et al. 2012). This is an emblematic example of how genomics contributed to a better understanding of evolutionary processes and it further underlines the importance of distinguishing gene trees from species trees.

In chapter 3 we inferred the evolutionary history of *B. graminis* based on genomic data, and estimated the divergence time between lineages. We showed that some lineages diverged around 20 million years ago while others are much younger. This work hopefully provides a backbone for the phylogeny and the timing of the evolution of *B. graminis* lineages on which the community can agree on (Troch et al. 2014, Panstruga and Spanu 2014). Finally also the work on evolution of effectors presented in chapter 4 can be correctly interpreted only in the light of the neutral evolution of *B. graminis*, it would be otherwise impossible to identify patterns of gene evolution that deviate from or follow the neutral tree.

This PhD project yielded many new insights, but many questions are still unanswered. In particular the origin and host range of under-sampled lineages, especially the one infecting wild grasses, is poorly understood. This will require massive sampling and phenotyping efforts, and a complete reconstruction of the evolution of host ranges in *B. graminis* lineages is probably very difficult to achieve. In particular it will be interesting to sequence and analyze genomes of additional lineages (for example the so far unsampled *ff.spp. agropyiri* and *bromi*) and additional isolates of the same lineage (ideally this should be done for all lineages, but realistically it is achievable for *B.g. tritici*

and/or *B.g. hordei*) to have a more complete picture of the macro and micro evolutionary processes acting in *B. graminis*.

Host specificity in *B. graminis*

Some of the work performed in my PhD has been on the host specificity of *B. graminis*. In particular the infection tests described in chapter 2 involved three *ff.spp.* (*triticale, tritici, secalis*) and their host(s). We found that some isolates of the *ff.spp. triticale* and *tritici* can grow to a very limited extent on the non-host species rye. This finding supports the notion that host specialization in *B. graminis* is not absolute (Troch et al. 2014). However no growth was observed after infection of distantly related species (barley, Poa, Dactylis, Lolium) with the *ff.spp. tritici*, *triticale* and *secalis*. This is in contrast with the observations of Eshed and Wahl (1970) and Sheng et al. (1993 and 1995) who found that isolates of the *ff.spp. tritici*, *hordei* and *avenae* could infect several wild grasses, and supports the finding of Wyand and Brown (2003) which also observed a stronger host specialization. As we used only few grass species and only one ecotype per species, ideally future studies aiming to understand the host specificities of *B. graminis* should use a large set of isolates to infect several varieties and ecotypes of different species. This can be particularly challenging, especially for forms which grow on wild grasses which are difficult to sample and to propagate in laboratories. The maintenance of isolates of these forms in laboratory conditions is very challenging, time consuming and requires large growth spaces, therefore to perform extensive infection tests is impossible. For this reason we could not determine the host range of the *ff.spp. poae*, *dactylidis* and of the form infecting *Lolium*.

Another point that future studies will need to take into account is the effects of physiological conditions on the infection tests. We observed in some cases that different replicates of the same infection test can give different results. For example it appears that different sections of the rye leaf (basal and distal) are differently susceptible to the *B.g. triticale* powdery mildew. Moreover also the different leaves (first leaf and second leaf) show different levels of susceptibility (M. Müller personal communication). These differences are obviously not due to the genetics of pathogen and host and have to be controlled when doing host specificity studies.

The finding of the work presented in chapter 2 opened the possibility to study the genetic determinant(s) of the host specificity in *B.g. triticale*. In particular we found that two *ff.spp.*, *tritici* and *triticale* are inter-fertile, they can grow on a common host (therefore they can be crossed) and they have different host range (*B.g. triticale* can grow on wheat and triticale while *B.g. tritici* can grow only on wheat). This can be exploited to generate recombinant progenies which can be used in genetic mapping or quantitative trait mapping to identify gene(s) responsible for the extended host

range of *B.g triticale*. Another approach to identify this/these gene(s) is genome wide association analysis using a large population of natural isolates of both *ff.spp.* (*tritici and triticale)*.

More lineages of the same *f.sp.*, the enigmatic case of *B.g. dicocci*

In chapters 2 and 3 we described a lineage composed of six *B. graminis* isolates from Israel. Isolates belonging to this lineage are genomically different form all other *B. graminis* isolates, and the most closely related isolates belong to the wheat powdery mildew (*B.g. tritici*). The first phenotyping tests that we performed showed that these isolates can grow only on tetraploid (pasta) wheat and not on hexaploid (bread) wheat. In these tests we used only two varieties of bread wheat (Kanzler and Chancellor) which have been susceptible to all wheat powdery mildew strains ever tested in our laboratory and do not contain known resistance genes. However, Ben-David and colleagues (2016) extended the set of varieties tested by us and found some hexaploid wheat lines on which these isolates could grow. This means that all these isolates belong to the same *f.sp.*, *B.g. tritici* because they all grow at least on some varieties of both hexaploid and tetraploid wheat.
It is difficult to interpret the observed genomic differences between these two group of isolates in light of the new phenotypic data. These two groups of isolates co-occur in the same geographic region (Israel) and can grow on the same host (tetraploid and exaploid wheat), but they form two different lineages. This implies the presence of some barrier to gene flow. A similar situation is observed in barley powdery mildew (*B.g. hordei*), here it was found that isolates with the same host range can be divided in two groups based on genomic data (E. Kominkova personal communication). These diversity patterns in *B.g tritici* and *B.g hordei* could be explained by particular demographic histories where sub-populations where first isolated, got in contact again later but they could not mate efficiently any more. To test these hypothesis crosses between isolates belonging to the two groups have to be performed and the viability of the offspring has to be analyzed. Another possibility is that these two genomic groups are specialized on different varieties of wheat and barley with a different genetic backgrounds, and even if under laboratory conditions they can grow on the same host, in natural environments they proliferate on different plants and therefore the probability of mating with members of the other group is reduced. This hypothesis is difficult to test because is not clear how to simulate “sub-natural“ conditions in the laboratory. However, quantitative phenotyping based on macro or microscopic high-throughput imaging could help to quantify small differences in growth efficiency which are not detected by visual observation of the phenotyping plates. These phenotypic differences could then be correlated to the genomic clusters.

Evolutionary genomics for phytopathology

B. graminis is used in many genetic, phytopathological and molecular studies together with its hosts, in particular wheat and barley. The last years were rich in exciting developments in this field and a large part of them would have been impossible or much slower without the sequencing of the reference genomes of *B. graminis hordei* and *tritici* (Spanu et al. 2010 and Wicker et al. 2013). One of the major advances was the identification of several hundreds of fungal effectors and some of their molecular targets in the plant cell (Pedersen et al. 2012, Zhang et al. 2012, Wicker et al. 2013, Schmidt et al. 2014, Pennington et al. 2016). Another exciting finding was the identification of four avirulence genes in *B. graminis* which are recognized by NB-LRR plant resistance proteins (Bourras et al. 2015, Praz et al. 2016, Lu et al. 2016).

Dobzhansky (1973) said: *"Nothing in biology make sense except in the light of evolution"* meaning that all biological phenomena should be observed in the framework of evolution.

Because of this, I think that genomics, and also evolutionary genomics can contribute in the future to advances in phytopathology. Having a good knowledge of the neutral evolution of *B. graminis* is fundamental to predict/simulate the pattern of genetic diversity in a population. The identification of outliers that do not fit in this pattern could lead to the discovery of genes under selection which may be involved in host adaptation or virulence in general. Moreover the observed diversity of known effectors and avirulence genes can be correctly interpreted only in the light of the neutral evolution of *B. graminis*, for example it is known that particular demographic histories can have the same effect of selection on genetic diversity. It is therefore necessary to be able to distinguish between adaptive and neutral processes and only a proper evolutionary genomic analysis can approach this goal.

REFERENCES

2.0, G. PTREP12. at <http://wheat.pw.usda.gov/ITMI/Repeats/>

Abascal, F., Zardoya, R. & Telford, M. J. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res. 38, W7–13 (2010).

Akaike. H. A new look on the statistical model dentification. IEE Transaction on Automatic Control 19(6), 716-723 (1974).

Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402 (1997).

Arch, È. et al. BRI1 is a critical component of a plasma-membrane receptor for plant steroids. 410, 380–383 (2001).

Beimforde C., Feldberg K., Nylinder S., Rikkinen J., Tuovila H., Dürfelt H., Gube M., Jackson D.J., Reitner J., Seyfullah L.J., Schmidt A.R. 2014. Estimating the Phanerozoic history of the Ascomycota lineages: Combining fossil and molecular data. Mol. Phylogenet. Evol. 77:307–319.

Ben-David, R., Parks, R., Dinoor, A., Kosman, E., Wicker, T., Keller, B., & Cowger, C. (2016). Differentiation Among Blumeria graminis f. sp. tritici Isolates Originating from Wild Versus Domesticated Triticum Species in Israel. *Phytopathology*, PHYTO-07.

Birch, P. R. J. et al. Oomycete RXLR effectors: delivery, functional redundancy and durable disease resistance. Curr. Opin. Plant Biol. 11, 373–9 (2008).

Bouchenak-Khelladi Y., Salamin N., Savolainen V., Forest F., Bank M .Van Der, Chase M.W., Hodkinson T.R. 2008. Large multi-gene phylogenetic trees of the grasses (Poaceae): progress towards complete tribal and generic level sampling. Mol. Phylogenet. Evol. 47:488–505.

Bouchenak-Khelladi Y., Verboom G.A., Savolainen V., Hodkinson T.R. 2010. Biogeography of the grasses (Poaceae): A phylogenetic approach to reveal evolutionary history in geographical space and geological time. Bot. J. Linn. Soc. 162:543–557.

Bourras S., McNally K.E., Ben-David R., Parlange F., Roffler S., Praz C.R., Oberhaensli S., Menardo F., Stirnweis D., Frenkel Z., Schaefer L.K., Fluckiger S., Treier G., Herren G., Korol A.B., Wicker T., Keller B. 2015. Multiple Avirulence Loci and Allele-Specific Effector Recognition Control the Pm3 Race-Specific Resistance of Wheat to Powdery Mildew. Plant Cell 27:2991–3012.

Bozkurt, T. O., Schornack, S., Banfield, M. J. & Kamoun, S. Oomycetes, effectors, and all that jazz. Curr. Opin. Plant Biol. 15, 483–92 (2012).

Bradbury, P. J. et al. TASSEL: software for association mapping of complex traits in diverse sample. Bioinformatics 23(19), 2633-2635 (2007).

Brasier C.M., Cooke, D. E. L., Duncan, J. M. Origin of a new *Phytophthora* pathogen through interspecific hybridization. *Proc. Natl. Acad. Sci.* 96, 5878–5883 (1999).

Brasier, C. The rise of the hybrid fungi. *Nature* 405, 134–5 (2000).

Brasier, C. M., Kirk, S. A. Rapid emergence of hybrids between the two subspecies of *Ophiostoma novo-ulmi* with a high level of pathogenic fitness. Plant Pathol. 59, 186–199 (2010).

Braun, U. (2011). The current systematics and taxonomy of the powdery mildews (Erysiphales): an overview. *Mycoscience*, *52*(3), 210-212.

Brown, J. K. M. & Hovmøller. Aerial dispersal of pathogens on the global and continental scales and its impact on plant disease. Science 297, 537-541 (2002).

Bustos, A. De & Jouve, N. Phylogenetic relationships of the genus Secale based on the characterisation of rDNA ITS sequences. 235, 147–154 (2002).

Cantarel, B. L., Korf, I., Robb, S. M., Parra, G., Ross, E., Moore, B., ... & Yandell, M. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome research*, *18*(1), 188-196 (2008).

Catanzariti, A.-M., Dodds, P. N., Lawrence, G. J., Ayliffe, M. a & Ellis, J. G. Haustorially expressed secreted proteins from flax rust are highly enriched for avirulence elicitors. Plant Cell 18, 243–56 (2006).

Chaw S.M., Chang C.C., Chen H.L., Li W.H. 2004. Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. J. Mol. Evol. 58:424–441.

Coppin, E., Debuchy, R., Arnaise, S. & Picard, M. Mating types and sexual development in filamentous ascomycetes. Microbiol. Mol. Biol. Rev. 61, 411–428 (1997).

Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., DePristo M.A., Handsaker R.E., Lunter G., Marth G.T., Sherry S.T., McVean G., Durbin R. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

Darwin, C. (1859) *On the Origin of Species*, London: John Murray.

Dean R., Van Kan J.L., Pretorius Z., Hammond-Kosack K.E., Di Pietro A., Spanu P.D., Rudd J.J., Dickman M., Kahmann R., Ellis J., Foster G.D. 2012. The Top 10 fungal pathogens in molecular plant pathology. Mol. Plant Pathol. 13:414–430.

Degnan, J. H. & Salter, L. A. Gene tree distributions under the coalescent process. *Evolution* 59(1), 24-37 (2005).

de Jonge, R., Bolton, M. D. & Thomma, B. P. H. J. How filamentous pathogens co-opt plants: the ins and outs of fungal effectors. Curr. Opin. Plant Biol. 14, 400–6 (2011).

DePristo, M. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491-498 (2011).

Dobin, A. Et al.STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1), 15-21 (2013).

Dou, D. et al. RXLR-mediated entry of Phytophthora sojae effector Avr1b into soybean cells does not require pathogen-encoded machinery. Plant Cell 20, 1930–47 (2008).

Edgar R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Ellis, J. G. & Dodds, P. N. Showdown at the RXLR motif: Serious differences of opinion in how effector proteins from filamentous eukaryotic pathogens enter plant cells. Proc. Natl. Acad. Sci. U. S. A. 108, 14381– 2 (2011).

Enright A.J., Van Dongen S., Ouzounis C.A. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research 30(7):1575-1584 (2002).

Eshed, N. & Wahl, I. Host ranges and interrelations of Erysiphe graminis hordei, E. graminis tritici and E. graminis avenae. Phytopathology 60, 628-634 (1970).

Eshed N., Dinoor, A., & Litwin, Y. The physiological specialization of wheat powdery mildew in Israel and the search for mildew resistance in wild wheat *Triticum dicoccoides*. *Phytoparasitica* 22: 49-90 (1994).

Excoffier L., Dupanloup I., Huerta-Sanchez E., Sousa V.C., Foll M. 2013. Robust Demographic Inference from Genomic and SNP Data. PLoS Genet. 9(10): e1003905.

Farrer, R. A. et al. M. C. Multiple emergences of genetically diverse amphibian-infecting chytrids include a globalized hypervirulent recombinant lineage. Proc. Natl. Acad. Sci. 108, 18732–18736 (2011).

Feldman, M. & Levy, A. A. Genome evolution due to allopolyploidization in wheat. Genetics 192, 763–74 (2012).

Felsenstein, J. . Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Biology*, *22*(3), 240-249 (1973).

Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, *17*(6), 368-376 (1981)

Galagan J.E., Calvo S.E., Borkovich K.A., Selker E.U., Read N.D., Jaffe D., Fitzhugh W., Ma L., Smirnov S., Purcell S., Rehman B., Elkins T., Engels R., Wang S., Nielsen C.B., Butler J., Endrizzi M., Qui D., Ianakiev P., Ianakiev P, Bell-Pedersen D., Nelson M.A., Werner-Washburne M., Selitrennikoff C.P., Kinsey J.A., Braun E.L., Zelter A., Schulte U., Kothe G.O., Jedd G., Mewes W., Staben C., Marcotte E., Greenberg D., Roy A., Foley K., Naylor J., Stange-Thomann N., Barrett R., Gnerre S., Kamal M., Kamvysselis M., Mauceli E., Bielke C., Rudd S., Frishman D., Krystofova S., Rasmussen C., Metzenberg R.L., Perkins D.D., Kroken S., Cogoni C., Macino G., Catcheside D., Li W., Pratt R.J., Osmani S.A., DeSouza C.P., Glass L., Orbach M.J., Berglund J.A., Voelker R., Yarden O., Plamann M., Seiler S., Dunlap J., Radford A., Aramayo R., Natvig D.O., Alex L.A., Mannhaupt G., Ebbole D.J., Freitag M., Paulsen I., Sachs M.S., Lander E.S., Nusbaum C., Birren B. 2003. The genome sequence of the filamentous fungus *Neurospora crassa*. Nature 422:859–868.

Garrison., E. & Marth, G. Haplotype-based variant detection from short-read sequencing . arXiv preprint arXiv:1207.3907 (2012).

Giraldo, M. C. & Valent, B. Filamentous plant pathogen effectors in action. Nat. Rev. Microbiol. 11, 800–14 (2013).

Glawe, D. A. (2008). The powdery mildews: a review of the world's most familiar (yet poorly known) plant pathogens. *Phytopathology*, *46*(1), 27.

Godfrey, D. et al. Powdery mildew fungal effector candidates share N-terminal Y/F/WxC-motif. BMC Genomics 11, 317 (2010).

Goss, E. M. et al. The plant pathogen *Phytophthoraandina* emerged via hybridization of an unknown *Phytophthora* species and the Irish potato famine pathogen, *P. infestans*. PloS One 9, e24543 (2011).

Goss, E. M., Press, C. M. & Grünwald, N. J. Evolution of RXLR-class effectors in the oomycete plant pathogen Phytophthora ramorum. PLoS One 8, e79347 (2013).

Gould S.J. and Lewontin R. C. The spandrels of San Marco and the panglossian paradigm: a critique of the adaptationist programme. *Proceedings of the Royal Society of London B: Biological Sciences*, *205*(1161), 581-598 (1979).

Granado, J., Felix, G., Boller, T., Institut, B. & Basel, U. Perception of Fungal Sterols in PIants '. 485–490 (1995).

Hacquard S., Kracher B., Maekawa T., Vernaldi S., Schulze-Lefert P., Ver Loren van Themaat E. 2013. Mosaic genome structure of the barley powdery mildew pathogen and conservation of transcriptional programs in divergent hosts. Proc. Natl. Acad. Sci. U. S. A. 110:E2219–28.

Hardison, J. R. Specialization of pathogenicity in Erysiphe graminis on wild ad cultivated grasses Phytopathology 34, 1-20 (1943).

Hein, H., Schierup, M. H. & Wiuf, C. Genes genealogies, variation and evolution. A primer in coalescent theory. (Oxford University Press, 2005).

Hermansen J. E., Torp, U. & Prahm, L. P. Studies of transport of live spores of cereal mildew and rust fungi across the north sea. Grana 17(1), 41-46 (1978).

Hiura U. 1978. Genetic basis of formae speciales. In: The Powdery Mildew. Spencer, D.M. p. 101–128.

Hiura U. 1965. Sexual compatibility between form species of Erysiphe graminis DC, and variability of the ascospore derived from the inter-form-specific hybridization. Nogaku Kenkyu 51:67–73.

Huelsenbeck J.P., Larget B., Alfaro M.E. 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. Mol. Biol. Evol. 21:1123–33.

Huson, D. H. & Scornavacca, C. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. Syst. Biol. 0(0):1-7 (2012).

Inohara, N. & Nuñez, G. ML -- a conserved domain involved in innate immunity and lipid metabolism. Trends Biochem. Sci. 27, 219–21 (2002).

Inuma T., Khodaparast S.A., Takamatsu S. 2007. Multilocus phylogenetic analyses within Blumeria graminis, a powdery mildew fungus of cereals. Mol. Phylogenet. Evol. 44:741–51.

Jiang, R. H. Y., Tripathy, S., Govers, F. & Tyler, B. M. RXLR effector reservoir in two *Phytophthora* species is dominated by a single rapidly evolving superfamily with more than 700 members. Proc. Natl. Acad. Sci. U. S. A. 105, 4874–9 (2008).

Jombart, T. & Ahmed, I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data.

Bioinformatics doi: 10.1093/bioinformatics/btr521 (2011).

Jones, J. D. G. & Dangl, J. L. The plant immune system. Nature 444, 323–9 (2006).

Kale, S. D. et al. External lipid PI3P mediates entry of eukaryotic pathogen effectors into plant and animal host cells. Cell 142, 284–95 (2010).

Joshi N.A., Fass J.N. 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at https://github.com/najoshi/sickle.

Junier T., Zdobnov E.M. 2010. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics 26:1669–1670.

Khang, C. H., Park, S.-Y., Lee, Y.-H., Valent, B. & Kang, S. Genome organization and evolution of the AVR-Pita avirulence gene family in the *Magnaporthe grisea* species complex. Mol. Plant. Microbe. Interact. 21, 658–70 (2008).

Klemptner, R. L., Sherwood, J. S., Tugizimana, F., Dubery, I. a & Piater, L. a. Ergosterol, an orphan fungal microbe-associated molecular pattern (MAMP). Mol. Plant Pathol. 15, 747–61 (2014).

Koonin, E. V. (2009). Darwinian evolution in the light of genomics. *Nucleic acids research*, *37*(4), 1011-1034.

Kusch, S., Ahmadinejad, N., Panstruga, R., & Kuhn, H. (2014). In silico analysis of the core signaling proteome from the barley powdery mildew pathogen (*Blumeria graminis f. sp. hordei*). BMC genomics, 15(1), 1.

Langmead B., Salzberg S.L. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9:357–9.

Lemey P., Salemi M. & Vandamme A-M. The phylogenetic handbook. Cambridge university press (2009).

Lepage T., Bryant D., Philippe H., Lartillot N. 2007. A general comparison of relaxed molecular clock models. Mol. Biol. Evol. 24:2669–80.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–9.

Liao. Y., Smyth, G. K. & Shi, W. featureCounts : an efficient general-purpose program for assigning sequence reads to genomic features. Bioinformatics 30(7), 923-930 (2014).

Limpert, E., Godet, F., Müller, K. Dispersal of cereal mildews across Europe. Agricultural and Forest Meterology 97, 293-308 (1999).

Lipka A.E., Tian F., Wang Q., Peiffer J., Li M., Bradbury P.J., Gore M., Buckler E.S., Zhang Z. 2012. GAPIT: genome association and prediction integrated tool. Bioinformatics 28:2397–9.

Liu, N. et al. Over-summering of wheat powdery mildew in Sichuan province, China. Crop Protection 34, 112–118 (2012).

Liu, X., Han, S., Wang, Z., Gelernter, J., & Yang, B. Z. Variant callers for next-generation sequencing data: a comparison study. *PLoS One*, *8*(9), e75619 (2013).

Luig, N. H. & Watson, I. A. The role of wild and cultivated grasses in the hybridization of f*ormae speciales* of *Puccinia graminis*. Aust. J. biol. Sci. 25, 335-342 (1972).

Ma, W., & Guttman, D. S. Evolution of prokaryotic and eukaryotic virulence effectors. *Current opinion in plant biology*, *11*(4), 412-419 (2008).

Maddison, W. P. (1997). Gene trees in species trees. *Systematic biology*, *46*(3), 523-536.

Maddison, W. P. & Knowles, L. L. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55(1) 21–30 (2006).

Marchal E. 1902. De la specialisation du paristisme chez l'Erysiphe graminis. Acad. Sci. Paris 135:210–212.

Marchler-Bauer, A. et al. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res. 39, D225–9 (2011).

Mascher, F., Reichmann, P. & Schori, A. Impact de l'oïdium sur la culture du triticale. *Revue Suisse Agric.* 38(4), 193–196 (2005).

Menardo F., Praz C., Wyder S., Bourras S., McNally K.E., Parlange F., Riba A., Roffler S., Schaefer L., Shimizu K.K., Valenti L., Zbinden H., Wicker T., Keller B. 2015. Hybridization of powdery mildew strains gives raise to pathogens on novel agricultural crop species. Nat. Genet. 48:1–24.

Middleton C.P., Senerchia N., Stein N., Akhunov E.D., Keller B., Wicker T., Kilian B. 2014. Sequencing of chloroplast genomes from wheat, barley, rye and their relatives provides a detailed insight into the evolution of the Triticeae tribe. PLoS One 9:e85761.

Miyake, K. Innate immune sensing of pathogens and danger signals by cell surface Toll-like receptors. Semin. Immunol. 19, 3–10 (2007).

Nater A., Burri R., Kawakami T., Smeds L., Ellegren H. 2015. Resolving evolutionary relationships in closely related species with whole-genome sequencing data. Syst. Biol. 64:1000–1017.

Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. 76, 5269–5273 (1979).

Newcombe, G., Stirling, B., McDonald, S. & Bradshaw, H. D. *Melampsora* × *columbiana*, a natural hybrid of *M. medusae* and *M. occidentalis. Mycol. Res.* 104, 261–274 (2000).

Nielsen, R. & Yang, Z. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148, 929–36 (1998).

Nowara, D. et al. HIGS: host-induced gene silencing in the obligate biotrophic fungal pathogen Blumeria graminis. Plant Cell 22, 3130–41 (2010).

Oberhaensli, S. et al. Comparative sequence analysis of wheat and barley powdery mildew fungi reveals gene collinearity, dates divergence and indicates host–pathogen co-evolution. Fungal Genet. Biol. 48, 327–334 (2011).

Oettler, G. The fortune of a botanical curiosity – Triticale: past, present and future. *J. Agric. Sci*. 143, 329 (2005).

Oku, T., Yamashita, S., Doi & Y, Hiura, U. Genetic Analysis of resistance Resistance of wheat cultivars to races and some *formae speciales* of *Erysiphe graminis* DC. Ann. Phytopath. Soc. Japan 52, 700–708 (1986).

O'Rawe, J. et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome med*, *5*(3), 28 (2013).

Pabinger, S. et al. A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in bioinformatics*, *15*(2), 256-278 (2014).

Panstruga R., Spanu P.D. 2014. Powdery mildew genomes reloaded. New Phytol. 202:13–14.

Parlange, F. et al. A mayor invasion of transposable elements accounts for the large size of the *Blumeria graminis* f.sp. *tritici* genome. Funct. Integr. Genomics, 11, 671-677 (2011).

Pendleton, A. L. et al. Duplications and losses in gene families of rust pathogens highlight putative effectors. Front. Plant Sci. 5, 299 (2014).

Pennington, H. G., Gheorghe, D. M., Damerum, A., Pliego, C., Spanu, P. D., Cramer, R., & Bindschedler, L. V. Interactions between the powdery mildew effector BEC1054 and barley proteins identify candidate host targets. *Journal of proteome research*, *15*(3), 826-839 (2016).

Perfect, S. E. & Green, J. R. Infection structures of biotrophic and hemibiotrophic fungal plant pathogens. Mol. Plant Pathol. 2, 101–8 (2001).

Pedersen, C., van Themaat, E. V. L., McGuffin, L. J., Abbott, J. C., Burgis, T. A., Barton, G., ... & Cramer, R. (2012). Structure and evolution of barley powdery mildew effector candidates. *Bmc Genomics*, *13*(1), 1.

Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat. Methods 8, 785–6 (2011).

Pfeifer, B., Wittelsbürger, U., Ramos-Onsin, S. E. & Lercher, M. JpopGenome: an efficient swiss army knife for population genomic analyses in R. Mol. Biol. Evol. 31(7), 1929-1936 (2014).

Pirooznia, M. et al. Validation and assessment of variant calling pipelines for next-generation sequencing. *Human genomics*, *8*(1), 14(2014).

Pliego, C. et al. Host-induced gene silencing in barley powdery mildew reveals a class of ribonuclease-like effectors. Mol. Plant. Microbe. Interact. 26, 633–42 (2013).

Posada D. 2016. Phylogenomics for Systematic Biology. Syst. Biol. 65:353–356.

Prieto M., Wedin M. 2013. Dating the diversification of the major lineages of Ascomycota (Fungi). PLoS One 8:e65576.

Pritchard, J. K., Stephens, M. & Donnerly, P. Inference of population structure using multilocus genotype data. Genetics 155, 945-959 (2000).

Queiroz K. De. 2007. Species concepts and species delimitation. Syst. Bot. 56:879–886.

Rafiqi, M., Ellis, J. G., Ludowici, V. a, Hardham, A. R. & Dodds, P. N. Challenges and progress towards understanding the role of effectors in plant-fungal interactions. Curr. Opin. Plant Biol. 15, 477–82 (2012).

Rannala B., Yang Z.H. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645–1656.

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Ridout, C. J. et al. Multiple avirulence paralogues in cereal powdery mildew fungi may contribute to parasite fitness and defeat of plant resistance. Plant Cell 18, 2402–14 (2006).

Robinson, M. D., McCarthy, D. J. and Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1), 139-140 (2010).

Ronquist F., Teslenko M., van der Mark P., Ayres D.L., Darling A., Höhna S., Larget B., Liu L., Suchard M., Huelsenbeck J.P. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61:539–42.

Rosenberg N.A., Nordborg M. 2002. Genealogical Trees, Coalescent Theory and the Analysis of Genetic Polymorphisms. Nat. Rev. Genet. 3:380–390.

Rossard, S., Roblin, G. & Atanassova, R. Ergosterol triggers characteristic elicitation steps in Beta vulgaris leaf tissues. J. Exp. Bot. 61, 1807–16 (2010).

Salamini, F., Ozkan, H., Brandolini, A., Schäfer-Pregl, R. & Martin, W. Genetics and geography of wild cereal domestication in the near east. *Nat. Rev. Genet*. 3, 429–41 (2002).

Sanabria, N. M., van Heerden, H. & Dubery, I. a. Molecular characterisation and regulation of a Nicotiana tabacum S-domain receptor-like kinase gene induced during an early rapid response to lipopolysaccharides. Gene 501, 39–48 (2012).

Saunders, D. G. O. et al. Using hierarchical clustering of secreted protein families to classify and rank candidate effectors of rust fungi. PLoS One 7, e29847 (2012).

Schmidt, S. M. et al. Interaction of a Blumeria graminis f. sp. hordei effector candidate with a barley ARF-GAP suggests that host vesicle trafficking is a fungal pathogenicity target. Mol. Plant Pathol. 15, 535–49 (2014).

Schulze-Lefert P., Panstruga R. 2011. A molecular evolutionary concept connecting nonhost resistance, pathogen host range, and pathogen speciation. Trends Plant Sci. 16:117–25.

Schwarz, G. E. Estimating the dimension of a model. Annals of Statistics 6(2), 461-464 (1978).

Shimizu-Inatsugi, R. et al. The allopolyploid *Arabidopsis kamchatica* originated from multiple individuals of *Arabidopsis lyrata* and *Arabidopsis halleri.* Molecular Ecology 18, 4024-4048 (2009).

Sota T. Radiation and reticulation: extensive introgressive hybridization in the carabid beetles *Ohomopterus* inferred from mitochondrial gene genealogy. *Popul. Ecol*. 44, 145-156 (2002).

Sousa V., Hey J. Understanding the origin of species with genome-scale data: modelling gene flow. Nat. Rev. Genet. 14:404–414 (2013).

Spanu P.D., Abbott J.C., Amselem J., Burgis T.A., Soanes D.M., Stüber K., van Themaat E.V.L., Brown J.K.M., Butcher S.A., Gurr S.J., Lebrun M.H., Ridout C.J., Schulze-Lefert P., Talbot N.J., Ahmadinejad N., Ametz C., Barton G.R., Benjdia M., Bidzinski P., Bindschedler L.V., Both M., Brewer M.T., Cadle-Davidson L., Cadle-Davidson M.M., Collemare J., Cramer R., Frenkel O., Godfrey D., Harriman J., Hoede C., King B.C., Klages S., Kleemann J., Knoll D., Koti P.S., Kreplak J., López-Ruiz F.J., Lu X., Maekawa T., Mahanil S., Micali C., Milgroom M.G., Montana G., Noir S., O'Connell R.J., Oberhaensli S., Parlange F., Pedersen C., Quesneville H., Reinhardt R., Rott M., Sacristán S., Schmidt S.M., Schön M., Skamnioti P., Sommer H., Stephens A,. Takahara H., Thordal-Christensen H., Vigouroux M., Wessling R., Wicker T., Panstruga R. et al. 2010. Genome Expansion and Gene Loss in Powdery Mildew Fungi Reveal Tradeoffs in Extreme Parasitism. Science 330:1543–1546.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Stergiopoulos, I. & de Wit, P. J. G. M. Fungal effector proteins. Annu. Rev. Phytopathol. 47, 233–63 (2009).

Stukenbrock, E. H. et al. The making of a new pathogen: insights from comparative population genomics of the domesticated wheat pathogen *Mycosphaerella graminicola* and its wild sister species. Genome Res. 21, 2157– 2166 (2011).

Stukenbrock, H. E., Christiansen, B. F., Hansen, T. T., Duteil, Y. J & Schierup, M. H. Fusion of two divergent fungal individuals led to the recent emergence of a unique widespread pathogen species. *Proc. Natl. Acad. Sci.* 27, 10954-10959 (2012).

Takamatsu, S. and Matsuda, S. Estimation of molecular clocks for ITS and 28s rDNA in Erysiphales. Mycoscience, 45, 340–344 (2004).

Takamatsu, S. Phylogeny and evolution of the powdery mildew fungi (Erysiphales, Ascomycota) inferred from nuclear ribosomal DNA sequences. Mycoscience 45, 147–157 (2004).

Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences.Lectures on Mathematics in the Life Sciences (American Mathematical Society).17, 57-86 (1986).

Than, C., Ruths, D. &Nakhleh, L. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics* 9:322 (2008).

Thompson, J. D., Higgins, D. G., & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, *22*(22), 4673-4680 (1994).

Tian, S., Weinert, J., Wolf, G.A. Infection of triticale cultivars by *Puccinia striiformis* : first report on disease severity and yield loss. J. of P. Dis. and Prot. 461–464 (2004).

Tosa, Y. Genetic analysis of the avirulence of wheatgrass powdery mildew fungus on common wheat. Genome, 32, 913–917. (1989a).

Tosa, Y. Evidence on wheat for gene-for-gene relationship between formae speciales of Erysiphe graminis and genera of gramineous plants. Genome, 32, 918– 924 (1989b).

Trinkaus, J. P., Biezen, E. A. Van Der & Jones, J. D. G. Curiosity and a dash of good luck The NB-ARC do main : a novel signalling motif shared by plant resistance gene products and regulators of cell death in animals. 225–227 (1980).

Troch, V., Audenaert, K., Bekaert, B., Höfte, M. & Haesaert, G. Phylogeography and virulence structure of the powdery mildew population on its “new” host triticale. *BMC Evol. Biol*. 12, 76 (2012).

Troch, V. et al.. Evaluation of resistance to powdery mildew in triticale seedlings and adult plants. *Plant Disease* 97, 410–417 (2013).

Troch V., Audenaert K., Wyand R., Haesaert G., Höfte M., Brown J.K.M. 2014. Formae speciales of cereal powdery mildew: close or distant relatives? Mol. Plant Pathol. 15:304–14.

Tyler, B. M. et al. Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis. Science 313, 1261–6 (2006).

Van der Biezen E.A., and Jones J.D. Plant disease-resistance proteins and the gene-for-gene concept. Trends Biochem. Sci. 23(12):454-456 (1998).

Walker, A. S., Bouguennec, A., Confais, J., Morgant, G. & Leroux, P. Evidence of host-range expansion from new powdery mildew (*Blumeria graminis*) infections of triticale (×*Triticosecale*) in France. *Plant Pathol*. 60, 207–220 (2011).

Weadick, C. J., & Chang, B. S. (2012). An improved likelihood ratio test for detecting site-specific functional divergence among clades of protein-coding genes. *Molecular biology and evolution*, *29*(5), 1297-1300.

Weßling, R., Epple, P., Altmann, S., He, Y., Yang, L., Henz, S. R., ... & Mukhtar, M. S. Convergent targeting of a common host protein-network by pathogen effectors from three kingdoms of life. *Cell host & microbe*, *16*(3), 364-375 (2014).

Win, J. et al. Effector biology of plant-associated organisms: concepts and perspectives. Cold Spring Harb. Symp. Quant. Biol. 77, 235–47 (2012).

Win, J. et al. Adaptive evolution has targeted the C-terminal domain of the RXLR effectors of plant pathogenic oomycetes. Plant Cell 19, 2349–69 (2007).

Wicker T., Oberhaensli S., Parlange F., Buchmann J.P., Shatalina M., Roffler S., Ben-David R., Doležel J., Simková H., Schulze-Lefert P., Spanu P.D., Bruggmann R., Amselem J., Quesneville H., Ver Loren van Themaat E., Paape T., Shimizu K.K., Keller B. 2013. The wheat powdery mildew genome shows the unique evolution of an obligate biotroph. Nat. Genet. 45:1092–1096.

Wu T.D., Nacu S. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26:873–881.

Wyand R., Brown J.K.M. 2003. Genetic and forma specialis diversity in Blumeria graminis of cereals and its implications for host-pathogen co-evolution. Mol. Plant Pathol. 4:187–98.

Xu, J., Vilgalys, R & Mitchell, T. G. Multiple gene genealogies reveal recent dispersion and hybridization in the human pathogenic fungus *Cryptococcus neoformans*. *Molecular Ecology* 9, 1471-1481 (2000).

Yaeno, T. et al. Phosphatidylinositol monophosphate-binding interface in the oomycete RXLR effector AVR3a is required for its stability in host cells to modulate plant immunity. Proc. Natl. Acad. Sci. U. S. A. 108, 14682–7 (2011).

Yang Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J. Mol. Evol. 39:306–314 (1994).

Yang Z. Maximum-Likelihood Estimation of Phylogeny from DNA Sequences When Substitution Rates Differ over Sites. Mol. Biol. Evol. 1:1396–1401 (1993).

Yang, Z., Nielsen, R., Goldman, N. & Pedersen, a M. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155, 431–49 (2000).

Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–91 (2007).

Yang, Z., Wong, W. S. W. & Nielsen, R. Bayes empirical bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. 22, 1107–18 (2005).

Yu, Y., Than, C., Degnan, J. H. &Nakhleh, L. Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. *Syst. Biol*. 60(2), 138-149 (2011).

Yu, Y., Degnan, J. H. &Nakhleh, L. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLOS Genetics* 8(4), e1002660 (2012).

Zhang, W.J. et al. Interaction of barley powdery mildew effector candidate CSEP0055 with the defence protein PR17c. Mol. Plant Pathol. 13, 1110–9 (2012).

Zhang, Z. et al. Of genes and genomes, needles and haystacks: Blumeria graminis and functionality. Mol. Plant. Pathol. 6, 561–575 (2005).

CURRICULUM VITAE

Surname:	Menardo
Name:	Fabrizio
Date of birth:	14.08.1987
Place of origin:	Cuneo, Italy

Education:

2002 – 2006	Maturità scientifica. Liceo Scientifico G. Peano, Cuneo Italy
2007 – 2010	Undergraduate studies in Natural Sciences. University of Torino
2010 – 2012	MSc in environmental Biology. University of Torino
2013 – present	PhD in Evolutionary Biology. University of Zurich

List of publications produced during the PhD

Menardo F, Wicker T and Keller B. 2016. Reconstructing the evolutionary history of grass powdery mildew lineages (*Blumeria graminis*) at different evolutionary time scales with NGS data. Submitted to Genome Biology and Evolution.

Praz RC, Bourras S, Zeng, F, Sanchez JM, Menardo F, et al. 2016. AvrPm2 encodes an RNase-like avirulence effector which is conserved in the two different specialized forms of wheat and rye powdery mildew. New Phytologist (in press).

Menardo F, Praz CR, Wyder S, Ben-David R, Bourras S, Matsumae H, McNally KE, Parlange F, Riba A, Roffler S, Schaefer L, Shimizu KK, Valenti L, Zbinden H, Wicker T & Keller B. 2016. Hybridization of powdery mildew strains give rise to pathogens on novel agricultural crop species. Nature Genetics 48 (2), 201-205.

Bourras S, McNally KE, Ben-David R, Parlange F, Roffler S, Praz CR, Oberhaensli S, Menardo F, Stirnweis D, Frenkel Z, Schaefer LK, Flückiger S, Treier G, Herren G, Korol AB, Wicker T & Keller B. 2015. Multiple avirulence loci and allele-specific effector recognition control the Pm3 race-specific resistance to wheat powdery mildew. The Plant Cell 27 (10), 2991-3012.

Parlange F, Roffler S, Menardo F, Ben-David R, Bourras S, McNally KE, Oberhaensli S, Stirnweis D, Buchmann G, Wicker T & Keller B. 2015. Genetic and molecular characterization of a locus involved in avirulence of *Blumeria graminis f. sp. tritici* on wheat Pm3 resistance alleles. Fungal Genetics and Biology 82, 181-192.

Roffler S, Menardo F & Wicker T. 2015. The making of a genomic parasite-the Mothra family sheds light on the evolution of Helitrons in plants. Mobile DNA 6 (1), 1.

ACKNOWLEDGMENTS

My PhD project was possible and successful only because of the contribution of many people. I would like to thank Beat and Thomas for giving me this project and the possibility to do a PhD. Thanks to Prof. Kentaro Shimizu for being part of my committee. Thanks to all collaborators, colleagues and friends for their help and presence. Finally, many thanks to my family, in particular my parents, my sister and Luisa for their constant love and support.